A new framework for structured variable selection and its application to Cox models with time-dependent covariates

Guanbo Wang

Doctor of Philosophy

Department of Epidemiology, Biostatistics and Occupational Health

McGill University
Montréal, Québec
April 2022

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

Dedication

I dedicate this thesis to the ones who loved, have loved, and will love me.

Acknowledgements

First and foremost, I would like to express my most genuine gratitude to my Ph.D. supervisors Drs. Mireille E. Schnitzer and Robert W. Platt. I am very pleased to be able to thank Mireille for her guidance (both in personal and career-oriented), support, caring and tolerance since I was a master student. She have educated me by showing her passion, rigorousness, persistence, resilience and smartness. Hundreds of insightful conversations with her have cultivated my statistical sense, her endless encouragement to pursue greater ambitions and confidence in me have given me the strength to continue on this demanding path. Her experience has encouraged me to further pursue an academic career. After I decided so, she started to train me as a young statistician. It is impossible for me to finish the journey without her. I am deeply grateful to Robert, for his support, guidance, generosity, optimism, flexibility, patience and valuable ideas. His advice definitely save me from making detours, and his endorsement boosts my confidence. The opportunities that he provided are highly appreciated. I have also benefited a lot from his philosophy of life. His backup as always makes me feel myself safe and strong. I would also like to express great thanks to both of my supervisors for financially supporting during this journey, and allowing me to explore whatever I like. I see both of them as a life-long advisors.

Thanks to the openness of my supervisors, I had the opportunity to visit Dr. Rui Wang at Harvard University in 2019, which undoubtedly hugely impacted my Ph.D. journey. Rui has been supportive as a family. I have benefited from her wisdom in life and her wide scope of knowledge in statistics. Her hardworking, persistence, and the notion of always doing high quality research have encouraged me a lot, and her selfless donation to my research made my journey smoother.

At the Department of Epidemiology, Biostatistics and Occupational Health, I am indebted to the several Professors who generously shared their knowledge and excitement for biostatistics

with me; in particular, I wish to thank Drs. Andrea Benedetti (who supervised me when I was a master student, and gave me many good advice), James Hanley, Erica Moodie, and Sahir Bhatnagar, for their teaching and mentoring. Their passion to Biostatistics also encouraged me to be a biostatistician. I am also very grateful to the departmental coordinators Andre Yves Gagnon and Katherine Hayden, for their support and for facilitating my studies at McGill University. I also wish to thank the Professors of the Department of Statistics at McGill University in particular Dr. Yi Yang. The data (which inspired all my Ph.D. work) that I analyzed were provided by Dr. Sylvie Perreault in Faculty of Pharmacy at University of Montreal, thank you for your financial support and contribution from your domain knowledge to my Ph.D. work. Thanks to Marc Dorais, who prepared the data, which saved me a lot of time. I also obtained support from Dr. Shu Yang in Statistics Department at North Carolina State University, though the work is not included in this thesis, her contribution is well acknowledged.

This work would not have been possible without the financial support of the Fonds de Recherche du Québec santé (FRQS), Research Institute of the McGill University Health Centre (RI-MUHC), Desjardins, Laboratoire de statistique – Centre de Recherches Mathématiques (CRM), and funding from Department of Epidemiology, Biostatistics and Occupational Health and Faculty of Medicine at McGill University. I acknowledge the usage of Compute Canada, an invaluable resource that enabled the computation in my studies.

I would like to express my gratitude to my internship mentors at Roche Dr. Artemis Koukounari and Melanie Poulin-Costello, for their confidence in me, support, openness, and encouragement.

I want to thank my friends in academia who have always been supporting me. In particular, Dr. Menglan Pang, for her support in both personal and academic life, and her accompany through my good and bad times; Dr. Tom Chen, for his intriguing ideas, selfless help, accompany and friendship; Dr. Xiao Wu, for the daily communication, discussion and encouragement; Steve Ferreira, as my best friend in the department, for the great times we

had, for translating the thesis abstract, and copy-editing several chapters of the thesis; Alex Levis, for the brotherhood and support since I came to Canada; I could not mention all, but the ones who have supported me include but not limit to Dr. Kaiqiong Zhao, Shomoita Alam, Shouao Wang, Yan Liu, Dr. Dongdong Li, Dr. Shuangning Li, Dr. Yue Song, Dr. Shu Di, Dr. Xihao Li and Larry Han. Among the friends who have accompanied me throughout the journey, I want to give special thanks to Siya Shao, for the endless laughter and deep thoughts about the life. My life will be less joyful if without Jixuan Li, Yihan Xing, Chenyu Tian, Dr. Carino Gurjao, Dr. Mansoureh Hakimi, Sean McGrath, Minghui Xiong, Zhuhua Qiu, Xinjia Zhang, Pengxiang Cui, Jialin Li and Xueyang Bai.

I would also like to thank Montreal where gave me an amazing experience, and the people in Montreal whoever gave me remarkable memories, for the accompany and love. I am forever indebted to my family, especially my parents Hongguang Jia and Zhiyou Wang, for always providing me with the unconditional love, ceaseless support and continual encouragement. I feel strong whenever I think of you. I am honored to be your son. A special thanks to my aunt Lianrong Wang and my cousin Fanshu Jiao, who would like to listen to me and provide me with more information as needed. I am blessed to be part of the big family.

Preface

This manuscript-based thesis consists of six chapters: an introduction, an original literature review, two chapters that correspond to two different stand-alone manuscripts, a technical chapter and a conclusion. A complete bibliography is presented at the end of this thesis. Chapters 3, 4 and 5 are main research topic of this thesis, concerning the structured variable selection problem, which provide another perspective to variable selection problems in general. Chapter 3 is a pure theoretical work, Chapter 4 and 5 contain a real world data application and a simulation study respectively.

The introduction, the literature review, and Conclusion (Chapters 1 2 and 6) of this thesis were written entirely by Guanbo Wang (GW) following further revised by Mireille E. Schnitzer (MES) and Robert W. Platt (RWP).

The ideas of Chapters 3 were conceived by GW. Tom Chen (TC) and MES helped in polishing the ideas. The writing was completed mainly by GW, and MES made significant contribution in some theoretical work development and improving the readability. RWP and Rui Wang (RW) further provide valuable comments and revised the article. A special thanks is given to Alex William Levis for the useful discussion.

Chapter 4 was conceptualized by GW and MES, with discussion of RWP and Sylvie Perreault (SP). The data were prepared by Marc Dorais. GW conduct the methodological work, writing, programming, and analysis. Substantive refining the article was given by MES, RWP, SP and RW.

For the Chapter 5, GW and Yi Yang (YY) initiated the idea. GW performed the methodological work for the project including theoretical derivations and simulations. GW wrote the entire Chapter. YY and TC helped in programming, and MES, RWP and RW contributed to editing the Chapter.

Abstract

Variable selection plays an important role in statistical modeling and prediction. It can discriminate between variables that are critical to predicting the outcome and the noise variables which are irrelevant or redundant for the purpose. Thus the "best" subset of variables can be identified for prediction. In addition, in a high-dimensional setting where the sample size is less than the number of covariates, variable selection can circumvent the identifiability issue by removing noise variables, and construct a valid predictive model. In practice, researchers often have knowledge of the relationships among covariates. For instance, an interaction is obtained from the product of two or more other variables (main terms). Taking such relationships into account in the implementation of variable selection can help to identify the relevant variable subsets and thus improve the prediction accuracy. My doctoral thesis establishes a general framework for incorporating these known relationships into variable selection, broadening its utility in applications and extending it to more types of data.

In the first manuscript, I propose a novel framework by first introducing the mathematical language of expressing selection rules (dependencies among the selection of variables). Then, I show that the resulting combination of permissible sets of selected variables ("selection dictionary") can be derived. I also bridge the proposed framework to existing penalized regression by offering a condition that relates to the selection dictionary: a postulated grouping structure (i.e., how to group variables in penalized regression) respecting the imposed selection rule.

The second manuscript involves an application of the theory and methods developed in the first one. The aim is to identify predictors of major bleeding among hospitalized hypertensive patients using oral anticoagulants for atrial fibrillation, where adherence and drug-drug interactions are considered. I illustrate how to use the framework in practice and provide a roadmap of how to identify the grouping structure to respect some common selection

rules.

In the third manuscript, I focus on a versatile (in terms of respecting selection rules) penalized regression, the overlapping group Lasso, and extend it to be used in the Cox model with time-dependent covariates. Technical details are presented in a more straightforward way to reach a broader audience. Simulation studies show that the proposed method is able to handle complex selection rules with the use of the framework. Furthermore, it can better identify the variables whose coefficients are non-zero, and is associated with a lower mean squared error as compared to the non-structured variable selection method.

In summary, the proposed framework highlights the importance of incorporating a priori knowledge of relationships among covariates into variable selection, advances the development of variable selection, and extends the use of existing methods.

Abrégé

La sélection de variables joue un rôle important dans la modélisation statistique et la prédiction. Cette méthode peut discriminer les variables authentiques, essentielles pour prédire le résultat, des variables nuisibles, qui ne sont pas pertinentes ou redondantes à cet effet. Ainsi, le « meilleur » sous-ensemble de variables peut être identifié pour la prédiction. De plus, dans un contexte de dimensionalité élevée où la taille d'échantillon est inférieure au nombre de covariables, la sélection de variables peut contourner le problème d'identifiabilité en supprimant les variables nuisibles et construire un modèle prédictif valide. En pratique, les chercheurs ont souvent une idée des relations entre les covariables. Par exemple, une variable (interaction) est obtenue à partir du produit de plusieurs autres variables (termes principaux). La prise en compte de ces relations dans la procédure de sélection de variables est utile afin de reconnaître les variables authentiques et ainsi améliorer la précision de la prédiction. Dans ma thèse doctorale, je me consacre à l'établissement d'un cadre général afin d'incorporer ces relations dans la sélection de variables, en élargissant son utilité aux applications et en le proposant pour d'autres types de données.

Dans le premier manuscrit, je propose le nouveau cadre en introduisant d'abord le langage mathématique d'expression des règles de sélection (dépendances entre la sélection des variables). Ensuite, la combinaison résultante d'ensembles autorisés de variables sélectionnées ("dictionnaire de sélection") peut être dérivée. Je relie également le cadre à la régression pénalisée existante en proposant une condition relative au dictionnaire de sélection : une structure de regroupement postulée (c'est-à-dire, comment regrouper les variables dans la régression pénalisée) respectant la règle de sélection imposée.

Le deuxième manuscrit accompagne le premier. L'objectif est d'identifier les facteurs prédictifs d'hémorragie majeure parmi les patients hospitalisés hypertendus utilisant des anticoagulants oraux pour la fibrillation auriculaire, où l'adhésion et l'interaction médicamenteuse sont prises en compte. En cours de route, j'illustre comment utiliser le cadre en pratique

et fournit une feuille de route sur la façon d'identifier la structure de groupement afin de respecter certaines règles de sélection communes.

Dans le troisième manuscrit, je me concentre sur une régression pénalisée polyvalente (en termes du respect des règles de sélection), le Group Lasso de chevauchement, et la développe pour le modèle de Cox avec des covariables dépendant du temps. Les détails techniques sont présentés de manière plus simple pour un public plus large. Des études de simulation démontrent que notre méthode peut gérer des règles de sélection complexes avec l'utilisation du cadre. De plus, elle permet de mieux identifier les variables dont les coefficients sont non nuls et est associée à une erreur quadratique moyenne inférieure par rapport à la méthode de sélection de variables non structurée.

En résumé, le cadre proposé met en évidence l'importance d'incorporer une connaissance a priori des relations entre les covariables dans la sélection de variables, fait progresser le développement de la sélection des variables et élargi l'utilisation des méthodes existantes.

Table of contents

List of Tables

List of Figures

Abbreviations

AIC Akaike Information Criterion

BIC Bayesian Information Criterion

CCVS Consistency of Categorical Variable Selection

CSH Consistency of Strong Hierarchy

CV-E Averaged cross-validated errors

FAMILY Framework for Modeling Interactions with a Convex Penalty

FAR False Alarm Rate

GLINTERNET Group Lasso Interaction Network

GRESH Group Regularized Estimation under Structural Hierarchy

HIERNET Strong Hierarchical Lasso

HQIC Hannan-Quinn information criterion

JDR Joint Detection Rate

MCP Minimax Concave Penalty

MIO Mixed-Integer Optimization

MR Missing Rate

MSE Mean Squared Errors

RAMQ Régie de l'Assurance Maladie du Québec

RCI Refitted C Index

SAIL Sparse Additive Interaction Learning

SCAD Smoothly Clipped Absolute Deviations

SHIM Strong Heredity Interaction model

VANISH Variable selection using Adaptive Nonlinear Interaction Structures in High dimensions

Chapter 1

Introduction

This thesis is concerned with variable selection that incorporates *covariate structure*, which indicates selection dependencies among variables. I focus on two important challenges. The first is to establish a new framework for variable selection that can integrate *a priori* knowledge about covariate structure, followed by an application to illustrate the framework. The second is to build a tool based on penalized regressions that can select structured time-dependent covariates when the outcome is time-to-event.

Model selection is a broad concept in statistics. It aims at selecting a statistical model (which may be non-linear) that satisfies a given condition. (Linhart and Zucchini, 1986; Zucchini, 2000; Johnson and Omland, 2004) Variable selection is a sub-concept of model selection. Under some assumptions made for statistical models (for example, the expected outcome is linearly dependent on covariates), we have some candidate (linear) models. Variable selection can then optimizes some criteria over the space of candidate models. (Kuo and Mallick, 1998; George, 2000)

Variable selection can be useful in many scenarios. For instance, when the number of covariates is larger than the sample size, it is infeasible to fit a regression model. One solution is to select variables, where the number is less than the sample size, that are most individually

predictive of the outcome, and then proceed with the regression. Another scenario is when researchers desire to learn the *sparsity pattern* of the data (i.e. which variables have non-zero coefficients). When there are many covariates in the data, variable selection can assist practitioners in identifying the variables that are most predictive of or most associated with the outcome.

Both constructing a predictive model in high dimensional data and identifying predictors are important in practice, notably in health science (Greenland, 1989), social science, (Handorf et al., 2020; Wu et al., 2020) and engineering (Peres and Fogliatto, 2018). In real-data applications, practitioners often have a priori knowledge of the covariate structures. For example, an interaction is derived from the product of main terms. As a second example, a categorical variable is typically represented by several dummy variables in the regression. See Chapter 2.1.2 for more examples. In variable selection, we can take this covariate structure into account by restricting the permissible variable combinations in the selected model to be compatible with the covariate structure. We refer to such restrictions as *selection rules*. For instance, if an interaction is selected, then main terms must be selected or as another example, dummy variables representing a categorical variable should all be selected collectively.

Integrating selection rules into variable selection can bring at least two benefits. First, it can ensure the interpretability of the selected model. For example, in predictor identification, if the interaction is selected without main terms, the interpretation of the coefficient would be different from the interpretation in the presence of the main terms. Second, variable selection techniques that respect selection rules have a higher chance of recovering the true sparsity pattern. (Yuan and Lin, 2006; Bhatnagar et al., 2020)

There is a substantial literature in the field of penalized regression and variable selection see Fan and Lv 2010 and references therein. Notably, group Lasso and its variations (Yuan and Lin, 2006; Wang and Leng, 2008; Jacob et al., 2009; Friedman et al., 2010b; Mairal et al.,

2010; Jenatton et al., 2011a; Bach et al., 2012a) can respect a certain types of selection rules, but not all. How to impose an arbitrary set of selection rules in variable selection is still an open question.

In this thesis, based on the identified research gap, I establish a new framework for variable selection that can respect any selection rule, and enrich some of the current solutions by extending it to accommodating survival outcomes and time-dependent covariates. In Chapter 2, I provide a detailed review of the motivation and current literature. In particular, I discuss variable selection recognizing covariate structures, penalized regression, and Cox models with time-dependent covariates.

In Chapter 3, I illustrate the new framework and its application to penalized regression and development of new variable selection methods. New definitions of mathematical objects are introduced, and their properties are investigated. The completeness and usefulness of the framework is also discussed.

In Chapter 4, I use an application in pharmacoepidemiology to demonstrate the utility of the framework and extend it by providing roadmaps of how to group variables in penalized regression for some common selection rules. The data used were compiled from a subset of the Régie de l'Assurance Maladie du Québec (RAMQ) drug and medical services database linked to the Med-Echo hospitalization database using encrypted patient healthcare insurance numbers. (Tamblyn et al., 1995; Wilchesky et al., 2004; Eguale et al., 2010; Perreault et al., 2020) The goal of the study is to identify predictors of major bleeding among hospitalized hypertensive patients using oral anticoagulants for atrial fibrillation.

In Chapter 5, I extend a versatile variable selection technique, developed by Mairal et al. (2010), to the Cox model with time-dependent covariates. Some challenges encountered in this setting are investigated, including algorithm design and computational burden. A simulation is conducted to show the advantages of the proposed method.

Chapters 3 and 4 were written as stand-alone manuscripts. In Chapter 6, I review the

significance of the research in this thesis, along with the limitations and future work.

Chapter 2

Literature review

In this Chapter, I review the variable selection problem in general, and specifically, address the importance of incorporating the covariate structure into variable selection. Then, I focus on penalized regression methods for variable selection, reviewing a few methods that can incorporate the covariate structure. Lastly, I review the Cox proportional hazards model in survival analysis and its extension to time-dependent covariates.

2.1 Variable selection incorporating covariate structures

2.1.1 The variable selection problem

Variable selection has drawn great attention as a statistical problem in recent decades. (Mehmood et al., 2012; Heinze et al., 2018) It is often referred to as the problem of selecting a subset of candidate explanatory variables (predictors) to predict the outcome.(Kuo and Mallick, 1998)

Variable selection can serve many goals in statistical analysis. First, it makes it possible to construct a predictive regression model with high dimensional data (where the sample size is less than the number of covariates) by selecting a limited number of predictors. Second, even

outside of the high-dimensional setting, removing spurious predictors in the predictive model can prevent the model from over-fitting and thus achieve a desirable prediction accuracy. Third, it can identify predictors among the candidate variables and thus provides both mechanistic insight of the prediction and improved prediction accuracy. (Heinze et al., 2018)

In general, variable selection addresses the scientific principle of parsimony, meaning that simpler models for describing reality are prioritized over complex ones. (Blumer et al., 1987; Seasholtz and Kowalski, 1993; Yarkoni and Westfall, 2017; Gauch Jr, 2002), In particular, variable reduction increases model utility and applicability, which improves accessibility for domain users such as clinicians.

Variable selection is often conducted under some model assumptions. For instance, it is commonly assumed that the outcome is linearly or non-linearly dependent on the predictors. Performing variable selection via non-parametric approaches or under non-linear assumptions (Smith and Kohn, 1996; Wang and Yin, 2008; Chung and Dunson, 2009; Genuer et al., 2010; Huang et al., 2010) may result in the selected model being less interpretable. In this thesis, I address goals related to retaining the interpretability of the selected model and thus focus on parametric specifications of the conditional mean outcome function.

"All subset selection" (Efroymson, 1960; Breaux, 1967) is one of the earliest developed model selection procedures. It was originally proposed as an exhaustive search method. Under the given model assumptions, with p covariates, it requires screening all possible (2^p) models based on a given criterion. Various of criteria can be considered, such as the greatest adjusted R^2; lowest Mallow's C_p (Mallows, 2000), lowest Akaike Information Criterion (AIC) (Akaike, 1998), AICc (Hurvich and Tsai, 1989), Bayesian Information Criterion (BIC), (Schwarz, 1978) or Hannan-Quinn information criterion (HQIC) (Hannan and Quinn, 1979), lowest cross-validated prediction error (Stone, 1974), etc. Stepwise regression can be implemented by the `R` package `leaps` (based on Fortran code by Alan Miller, 2020) for linear models.

However, this procedure is infeasible for a large p.

To overcome the large search space issue, backward and forward selection in stepwise regression were proposed. For instance, forward selection starts from a null model and adds one variable at a time to test its performance increment, where the performance is decided by a given criterion. Then the best move (addition of a given variable) can be determined. However, the methods were proved to be biased in estimation and inconsistent in selection; see (Hurvich and Tsai, 1990; Steyerberg et al., 1999; Whittingham et al., 2006; Doornik, 2009) for the proofs and discussion. Furthermore, when the covariates are of high dimension and no restriction is applied to the model space, such methods are infeasible to implement.

From the perspective of the bias-variance trade-off, one can fit a penalized regression where the coefficients are penalized toward 0, so that the model can achieve a better performance (for example, lower mean-squared error or prediction error at a price of a biased estimator). Hoerl and Kennard (1970) first introduced this concept by developing ridge regression. Later on Frank and Friedman (1993) developed bridge regression. Such methods were not designed directly to select variables, but the ideas generated further research on variable selection via penalized regression, which is introduced in section 2.2.

2.1.2 Covariate structures and selection rules

Covariate structures, or dependencies among the covariates, are often desirable depending on the data structure. They are not only beneficial when the goal is estimation, but also in variable selection.

In practice, there are different classes of covariate structures. Some examples are:

1. An interaction is the product of its main terms. Thus there are dependencies in how these terms are selected into the model.

2. Binary indicators (dummy variables) representing a categorical variable are correlated

to each other.

3. Tree-structured dependencies: for instance, in natural language processing, a topic is dependent on all of its sub-topics.

Incorporating those covariate structures into variable selection is beneficial, as doing so can better recover sparsity patterns and improve the prediction accuracy of the resulting models. More importantly, those dependencies are closely related to the interpretability of the resulting model. (Heinze et al., 2018; Thompson, 2009; Huang and Zhang, 2010; Obozinski et al., 2010) Failing to consider them may result in a model that is less interpretable, with a loss of mechanistic insight or conceptual understanding to the domain users. (Harrell Jr et al., 1984; Andersen and Bro, 2010)

One way to incorporate covariate structure into variable selection is to set selection rules. Examples of proper selection rules corresponding to the above examples are listed below.

1. If an interaction is selected, then all of its main terms must be selected, which is referred to as strong heredity, (Haris et al., 2016b; Bhatnagar et al., 2020) or, if an interaction is selected, then at least one of the main terms must be selected, which is referred to as weak heredity. (Yuan et al., 2009)

2. The dummy variables representing a categorical variable should be selected collectively, (Yuan and Lin, 2006) so that the categorical variable (e.g., race, age group) as a whole can be said to be predictive or not, rather than a single level. In addition, if they are not all selected, the reference group becomes heterogeneous.

3. In classifying online postings to newsgroups (for example, `alt.atheism` and `talk.religion.misc` (Lacoste-Julien et al., 2008; Zhu et al., 2012)), when covariates include both topics (e.g., sports) and its sub-topics (e.g., basketball), a selection rule can be set as "if the sub-topic is selected, then its topic should be selected."

We firstly propose and formalize the concept of selection rule in Chapter 3, and then develop

methods to express it in a systematic mathematical language.

As mentioned in section 2.1.1, it is often of interest to restrict the search space in stepwise and all-subsets regression. One byproduct of integrating the selection rule is the reduction of the number of candidate models while ensuring the model interpretability.

Much research has been done to incorporate selection rules into variable selection. They can be applied in many research fields, such as medical research (Wang et al., 2009), bioinformatics (Jacob et al., 2009; Kim and Xing, 2010), topic modeling (Jenatton et al., 2011b) and computer vision (Huang et al., 2011). Next, from the perspective of respecting selection rules, we review a selection of penalized regression techniques that are used in Chapters 3, 4 and 5.

2.2 Penalized regression for variable selection

Different techniques correspond to different types of selection rules. We use linear regression and its square loss as an example to illustrate these penalized regression methods for variable selection, but the methods have been extended to other model and loss function specifications.

Suppose the data consist of n observations and p covariates, and the covariates are centered about 0 (thus we omit the intercept). A linear model describes the following relationship

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{y} \in \mathbb{R}^n, \boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p\} \in \mathbb{R}^{n \times p}, \boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\epsilon} \in \mathbb{R}^n$ are the outcome, covariate matrix, coefficient vector, and i.i.d random errors respectively.

Penalized regression in general solves

$$\min_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) + \Omega(\boldsymbol{\beta}, \boldsymbol{\theta}), \tag{2.1}$$

where $\ell(\boldsymbol{\beta})$ is a convex loss function, in our case $\ell(\boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{X}) = \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{\beta}\boldsymbol{X}\|_2^2$ (which can also be generalized to weighted least-squares), $\boldsymbol{\theta}$ is a vector of hyper parameters, and the penalty term $\Omega(\boldsymbol{\beta}, \boldsymbol{\theta})$ is a function of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$.

The value of $\boldsymbol{\theta}$ controls the level of penalization. For example, in Lasso, a higher level of penalization corresponds to a smaller set of variables being selected. The shrinkage of coefficients usually improves the prediction accuracy because of the bias-variance trade-off. The value of $\boldsymbol{\theta}$ is often determined by cross-validation. Two criteria can be applied:

1. choose the $\boldsymbol{\theta}$ such that the cross-validated prediction error reaches its minimum value (min rule) (Hastie et al., 2009), or
2. choose the maximum $\boldsymbol{\theta}$ such that the cross-validated prediction error is within one-standard error of the errors. (1se rule) (Hastie et al., 2009; Breiman et al., 2017)

The choice of criteria should be based on the goal of conducting variable selection. (Meinshausen and Bühlmann, 2006) When the goal is to select a model with the highest prediction accuracy, the min rule should be used. When the goal is to recover the sparsity pattern, that is, select the most parsimonious model while maintaining the prediction accuracy, the 1se rule should be applied. (Hastie et al., 2009; Chen and Yang, 2021)

Different specifications of Ω result in different regularization methods, with different variable selection results. In general, the penalty term can be presented as a sum of norms of groups of coefficients of variables. Therefore, the specification of Ω always requires the specification of a grouping structure, which assigns different groups of coefficients of covariates into multiple norms.

A grouping structure $\mathbb{G} \coloneqq \{\mathbb{g}_i, i = 1, ..., I\}$ is a set of non-empty subsets of the candidate

variables whose union is the set of all candidate variables. Let $\boldsymbol{\beta}_{|\mathfrak{g}}$ be a vector of the same length as $\boldsymbol{\beta}$ whose coordinates are equal to those of $\boldsymbol{\beta}$ for indices in the set $\mathfrak{g}$ and 0 otherwise. We can thus replace $\Omega(\boldsymbol{\beta}, \boldsymbol{\theta})$ by $\Omega(\boldsymbol{\beta}_{|\mathfrak{g}}, \boldsymbol{\theta})$ in (2.1), so that we can use harmonized notation to introduce the following methods.

One can assess the properties of variable selection methods by the following criteria:

1. Selection consistency. An estimator is said to be consistent in selection if, as $n \to \infty$, each coefficient converges to a non-zero value when its true value is also non-zero, (Zhao and Yu, 2006; Yuan and Lin, 2007) and

2. The oracle property, which is a stronger property than selection consistency. The oracle property states that the asymptotic distribution of the estimator is the same as the asymptotic distribution of the maximum likelihood estimator on the true support. (Zou, 2006; Fan and Li, 2001) That is, the estimator performs as well as if the true model was known.

2.2.1 Best subset selection via optimization

Best subset selection, such as stepwise regression, is not implementable when the covariates are of high dimension, not only because of the computational burden, but also because, when $n < p$, it is infeasible to fit a regression model. Bertsimas et al. (2016) developed a computationally tractable method that converts best subset selection into a regularization problem, which constrains the number of variables being selected. It solves

$$\min_{\boldsymbol{\beta}} \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{\beta}\boldsymbol{X}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|_0 \leqslant k, \tag{2.2}$$

where $\|\boldsymbol{\beta}\|_0 = \sum_{i=1}^{p} \mathbb{1}(\beta_i \neq 0)$, and k is a tuning parameter. The problem can be reformulated as a mixed integer optimization problem (Bertsimas and Weismantel, 2005; Pochet and

Wolsey, 2006)

$$\min_{\boldsymbol{\beta}} \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{\beta}\boldsymbol{X}\|_2^2 \quad \text{s.t.} - Mz_i \leqslant \beta_i \leqslant Mz_i, z_i \in \{0,1\}, \sum_{i=1}^{p} z_i \leqslant k, i = 1, \ldots, p,$$

where M is a positive number that is no less than the maximum value in the solution vector of the above problem. Practically, it is chosen to be large enough to obtain the solution that equals the solution to (2.2). The proposed algorithm is based on projected gradient descent methods in first-order convex optimization (Nesterov, 2003) and was then generalized to the discrete optimization problem. (Bertsimas et al., 2016) The algorithm was proven to converge, but its selection consistency and oracle properties have not yet been established.

The current version of best subset selection cannot incorporate any selection rule, i.e., each covariate has a chance to be selected or not and it is not dependent on the selection of any other covariate. In Chapter 3, we propose some optimization problems that can integrate certain classes of selection rules.

2.2.2 Lasso

For the squared loss, Lasso (Least absolute shrinkage and selection operator) (Tibshirani, 1996) solves

$$\min_{\boldsymbol{\beta}} \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{\beta}\boldsymbol{X}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 , \tag{2.3}$$

where $\lambda > 0$ is a regularization parameter. Note that here there is only one hyper parameter λ, corresponding to $\boldsymbol{\theta}$ in (2.1). The L^1 penalty shrinks each coefficient individually towards 0 and the correlation among each covariates is not considered. No selection rule can be integrated into standard Lasso. The problem (2.3) is a convex optimization problem, which can be solved efficiently by least angle regression (Efron et al., 2004) or the coordinate

gradient descent algorithm (Friedman et al., 2007; Wu and Lange, 2008; Friedman et al., 2010a). Bertsimas et al. (2016) showed that Lasso is a weaker relaxation than best subset selection via optimization, in the sense that the minimum of problem (2.2) is lower-bounded by the optimum objective value of Lasso. Additionally, the algorithm for solving (2.2) performs better than Lasso in recovering the sparsity pattern.

Lasso suffers from a few drawbacks. It produces biased estimators for large coefficients (Fan and Li, 2001) and is consistent in selection only under certain conditions (Zou, 2006). Therefore, it does not enjoy the oracle property.

2.2.3 Adaptive Lasso

Several methods were developed to produce estimators possessing the oracle property. Adaptive Lasso (Zou, 2006) achieves the goal by a two-stage procedure. The adaptive Lasso solves

$$\min_{\boldsymbol{\beta}} \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{\beta}\boldsymbol{X}\|_2^2 + \lambda \sum_{j=1}^{p} \omega_j |\beta_j|, \tag{2.4}$$

where ω_j is a positive weight for each of the coefficients. The first stage of the procedure is to specify the weights. Denote the jth estimator from linear regression by $\hat{\beta}_j$. Practically, the weight ω_j can be $1/|\hat{\beta}_j|, j = 1, \ldots, p$. Then, the second stage is to fit the "weighted" Lasso. Intuitively, the two-stage procedure penalizes large coefficients less, resulting in an estimator with the oracle property. (Zou, 2006) The adaptive lasso can be solved as efficiently as Lasso using similar algorithms.

2.2.4 Smoothly Clipped Absolute Deviations (SCAD) and Minimax Concave Penalty (MCP)

While the adaptive Lasso reduces bias compared to Lasso by using a two-stage approach, single-stage methods can also achieve the same goal by setting nonconvex penalties, which have the advantage of penalizing less when the coefficients are larger. For example, denote Ω as $\Omega(\boldsymbol{\beta};\boldsymbol{\theta}) = \sum_{j=1}^{p} P(\beta_j;\lambda,\gamma)$. The SCAD penalty (Fan and Li, 2001) is defined as

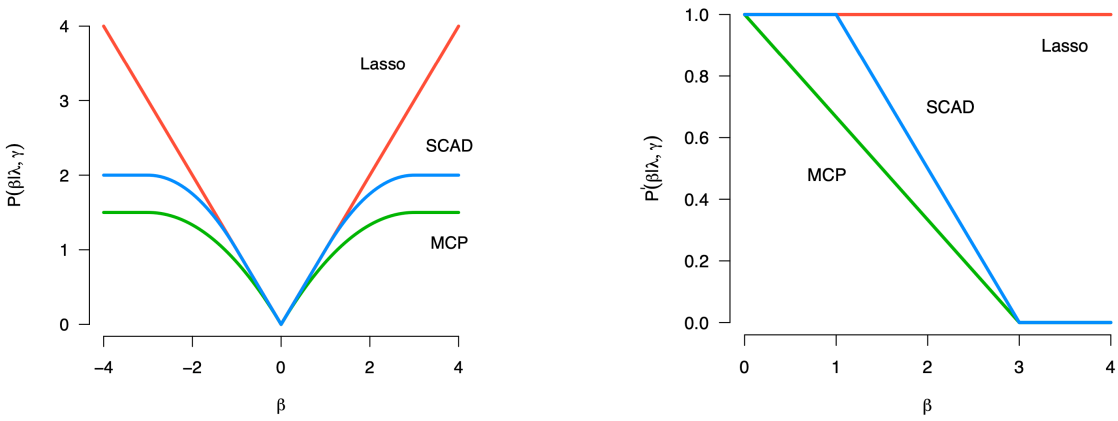
$$P(\beta_j;\lambda,\gamma) = \begin{cases} \lambda|\beta_j|, & \text{if } |\beta_j| \leqslant \lambda, \\ \frac{2\gamma\lambda|\beta_j| - \beta_j^2 - \lambda^2}{2(\gamma-1)}, & \text{if } \lambda < |\beta_j| < \gamma\lambda, \\ \frac{\lambda^2(\gamma+1)}{2}, & \text{if } |\beta_j| \geqslant \gamma\lambda, \end{cases} \tag{2.5}$$

where $\gamma > 2$, and the MCP is defined as

$$P(\beta_j;\lambda,\gamma) = \begin{cases} \lambda|\beta_j| - \frac{\beta_j^2}{2\gamma}, & \text{if } |\beta_j| \leqslant \gamma\lambda, \\ \frac{1}{2}\gamma\lambda^2, & \text{if } |\beta_j| > \gamma\lambda, \end{cases} \tag{2.6}$$

where $\gamma > 1$. Both λ and γ are tuning parameters. Such penalties are often referred to as folded concave penalties: it is concave on both the positive and negative sides of $|\beta_j|$ and also symmetric (or folded). We use Figure 2.1 (Breheny and Huang, 2015) to show how they reduce the bias as compared to Lasso when $\boldsymbol{\beta}$ is a single dimension coefficient. As we can see, with the increasing of the absolute values of the coefficient, the rate of penalization by Lasso does not change, whereas the SCAD and MCP level off, with the MCP penalizing less than the SCAD.

Both methods have the oracle property. The hyper parameter λ is chosen by cross-validation similar to Lasso and adaptive Lasso. Another hyper parameter γ, controls the concavity of the penalty. As $\gamma \to \infty$, both the SCAD and MCP converge to Lasso, and as γ approaches its minimum, the corresponding estimators are the least biased, but the estimates are unstable



(a) Penalty functions (b) Derivatives of penalty functions

Figure 2.1: Penalty functions and their derivatives of Lasso, SCAD and MCP (Breheny and Huang, 2015)

in the sense that there are multiple local minima in the optimization problems. (Breheny and Huang, 2015) In practice, γ is recommended to be 3.7 and 2.7 in the SCAD and MCP respectively. For more discussion on the choice of γ see (Mazumder et al., 2011; Breheny and Huang, 2011).

The nonconvex penalties can be approximated by local linear approximation (Zou and Li, 2008), and thus the objective functions of the SCAD and MCP can be optimized by the least angle regression algorithm. Alternatively, the nonconvex problem can also be solved by coordinate descent algorithms. (Breheny and Huang, 2011)

Though the methods described above cannot incorporate selection rules, they are the base of variable selection techniques. The ideas of bias reduction from the adaptive Lasso, the SCAD, and MCP can be utilized when penalizing the coefficients in a group fashion, which we introduce below.

2.2.5 Group Lasso and its variation

When the penalty term is defined as a function of $|\beta_j|, j = 1, \ldots, p$, the penalized regression selects variables individually. That is, the selection of each variable is independent from the

selection of others. Intuitively, if the penalty term is defined as a function of the norm of a group of coefficients, that is, $\left\|\boldsymbol{\beta}_{|\mathbb{g}}\right\|, \mathbb{g} \in \mathbb{G}$, then the penalized regression can select variables in a group fashion.

Notably, Yuan and Lin (2006) developed the group Lasso, which solves

$$\min_{\boldsymbol{\beta}} \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{\beta}\boldsymbol{X}\|_2^2 + \lambda \sum_{\mathbb{g} \in \mathbb{G}} \omega_{\mathbb{g}} \left\|\boldsymbol{\beta}_{|\mathbb{g}}\right\|_2,$$

where $\omega_{\mathbb{g}}$ is a positive weight for group $\mathbb{g}$. In practice, it is suggested to use $\omega_{\mathbb{g}} = \sqrt{|\mathbb{g}_j|}$. While the L^1 norm penalty of a vector produces sparsity among the coefficients in the vector, the L^2 norm penalty of a vector forces the coefficients in the vector to be selected collectively, i.e., they have to be selected/unselected together. The group Lasso was first proposed to be solved by the group least angle regression algorithm (Yuan and Lin, 2006) and was later proved to be solved more efficiently by the groupwise majorization descent algorithm (Yang and Zou, 2015; Lange et al., 2000).

Due to the similarity between Lasso and the group Lasso, the latter also lacks the oracle property. To overcome this issue, the adaptive group Lasso (Wang and Leng, 2008), the group SCAD (Wang et al., 2007,0; Breheny and Huang, 2015), and the group MCP (Huang et al., 2012; Breheny and Huang, 2015) were proposed, which replace the β_j by $\left\|\boldsymbol{\beta}_{|\mathbb{g}}\right\|_2$ in (2.4), (2.5) and (2.6) respectively. All three methods enjoy the oracle property.

The group Lasso and its variations can incorporate selection rules such as "select variables in a group $\mathbb{g}$ collectively". In other words, they select zero or all variables in a group of variables $\mathbb{g}$. However, these methods require that $\mathbb{G} = \{\mathbb{g}_i, i = 1 \ldots, I\}$ is a non-overlapping partition of the set of candidate covariates, meaning that the intersection of any two groups must be the empty set: $\mathbb{g}_i \cap \mathbb{g}_j = \emptyset, \forall \mathbb{g}_i \in \mathbb{G}, \forall \mathbb{g}_j \in \mathbb{G}, i \neq j$. This limits the incorporation of more complex selection rules.

2.2.6 Sparse group Lasso

The above methods create sparsity at a group level, but not at the level of individual variables, though this latter property is sometimes desired simultaneously. For example, in studying gene expression data, genes in a same pathway do not function independently in genetic conditions – some genes in a given pathway are collectively predictive of the outcome. The goal is to select both the predictive pathways and individual genes in the pathways. In this scenario, we would like to identify the active pathways in addition to selecting the driving genes indicated in the genetic condition or equivalently, identify the pathways and genes simultaneously.

Simon et al. (2013) developed the sparse group Lasso which is suitable for the above scientific and statistical goals. The corresponding penalty can be viewed as a convex combination of Lasso (L^1 penalty) and the group Lasso (L^2 penalty). The sparse group Lasso solves

$$\min_{\boldsymbol{\beta}} \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{\beta}\boldsymbol{X}\|_2^2 + (1-\alpha)\lambda \sum_{\mathbb{g}\in\mathbb{G}} \omega_{\mathbb{g}} \left\|\boldsymbol{\beta}_{|\mathbb{g}}\right\|_2 + \alpha\lambda \|\boldsymbol{\beta}\|_1 ,$$

where $\alpha \in [0,1]$, and $\mathbb{g}_i \cap \mathbb{g}_j = \emptyset, \forall \mathbb{g}_i \in \mathbb{G}, \forall \mathbb{g}_j \in \mathbb{G}, i \neq j$. With such a penalty, the selection of an individual variable is affected by both its own predictability and the predictability of the group that it belongs to. The authors proposed a blockwise descent algorithm to solve the optimization problem, where an accelerated gradient descent with backtracking line search is used to solve the estimators in each group. An improved version, called the adaptive sparse group Lasso (Poignard, 2016) (similar to the adaptive group Lasso) has the oracle property. Note that it is not the only technique (but may be the most popular one) that can handle such selection rules; the group bridge (Huang et al., 2009), the composite MCP (Breheny and Huang, 2009), and some methods introduced later can achieve the goal as well. See Huang et al. (2012) for a detailed review.

The above methods can respect selection rules such as "select groups of variables collectively,

and then select a number of variables in the selected groups", or equivalently, "select a number of variables in each of the selected groups". Note that the number of groups being selected and the number of variables being selected in each selected group cannot be pre-specified because they are entirely determined by the tuning parameters.

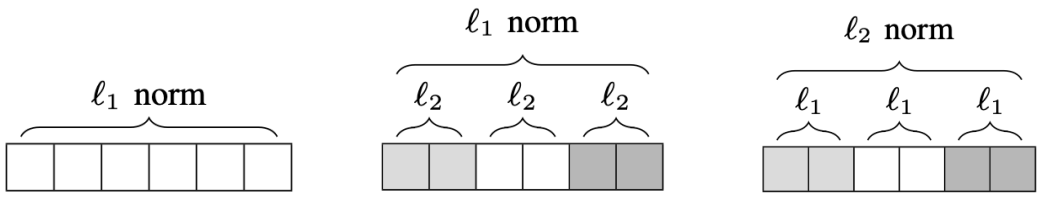
2.2.7 Exclusive (group) Lasso

With the above described data, researchers sometimes presume or know a priori that all the pathways are active, but only want to identify the representative genes in each pathway. In this case, it is still beneficial to consider the correlation among the variables in each group.

Zhou et al. (2010) developed the Exclusive Lasso, which solves

$$\min_{\boldsymbol{\beta}} \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{\beta}\boldsymbol{X}\|_2^2 + \lambda \sum_{\mathfrak{g}\in\mathbb{G}} \left\|\boldsymbol{\beta}_{|\mathfrak{g}}\right\|_1^2,$$

where $\mathfrak{g}_i \cap \mathfrak{g}_j = \emptyset, \forall \mathfrak{g}_i \in \mathbb{G}, \forall \mathfrak{g}_j \in \mathbb{G}, i \neq j$. Note that $\lambda_1 \sum_{\mathfrak{g}\in|\mathbb{G}|} \left\|\boldsymbol{\beta}_{|\mathfrak{g}}\right\|_1^2$ and $\lambda_2 \sqrt{\sum_{\mathfrak{g}\in|\mathbb{G}|} \left\|\boldsymbol{\beta}_{|\mathfrak{g}}\right\|_1^2}$ can produce equivalent sparsity with two different values of λ_1 and λ_2. (Bach et al., 2012b) In comparison with the sparse group Lasso, whose penalty is a convex linear combination of L^1 and L^2 norm penalties, the penalty in exclusive Lasso is called a composite of L^1 and L^2 norm penalties. (Campbell and Allen, 2017) To have a better understanding of the relationship among Lasso (L^1 norm penalty), the group Lasso ($L^{2,1}$ norm penalty), and the exclusive Lasso ($L^{1,2}$ norm penalty), we show the comparison of their penalty mechanisms in Figure 2.2 (Sun et al., 2020). Variables are represented by grids, and grids with the same color are specified in a same group. Lasso prompts sparsity at the individual variable level by introducing L^1 norm among the variables; the group Lasso enforces sparsity among the groups by applying L^1 norm to the groups, where the variables in each group are subject to L^2 penalty so that they must be selected collectively; the exclusive Lasso encourages sparsity among the variables in a same group through the L^1 penalty on the variables but, as a result



(a) Lasso (b) group Lasso (c) exclusive Lasso

Figure 2.2: A comparison of the penalties of Lasso, group Lasso and exclusive Lasso (Sun et al., 2020)

of introducing L^2 penalty into the groups, at least one variable in each of the groups must be selected.

Coordinate descent can fit the exclusive Lasso, and it was proved that the algorithm converges to the global minimum. (Sun et al., 2020)

While the exclusive Lasso requires the groups to be non-overlapping, Kong et al. (2014) developed the so-called exclusive group Lasso which relaxes this limitation. Multiple methods address this optimization problem, including an iteratively re-weighted algorithm (Kong et al., 2014; Yamada et al., 2017), a dual Newton based preconditioned proximal point algorithm (for a weighted version of the exclusive group Lasso) (Lin et al., 2019), and the active set algorithm (Gregoratti et al., 2021). They can converge to the global minimum with a fast convergence rate.

Under some conditions, the exclusive (group) Lasso is selection consistent, but its oracle property has not yet been established.

The exclusive (group) Lasso can respect selection rules of the type "select at least one variable in each (overlapped) group", which can restrict the number of variables being selected at a certain level (that is, at least one) in each group. If all the variables in a same group are correlated, then the method can be viewed as an approach to select uncorrelated variable sets. However, it still cannot respect selection rules of the type "selecting exactly c variables".

2.2.8 Overlapping group Lasso

In variable selection where interactions are of interest, because of the model interpretability and the principle of parsimony, common selection rules are strong heredity ("if an interaction is selected, then all the main terms must be selected"), and weak heredity ("if an interaction is selected, then at least one of the main terms must be selected"). (Yates, 1937; Hamada and Wu, 1992; Chipman, 1996; Joseph, 2006) Related research includes Strong Heredity Interaction model (SHIM) (Choi et al., 2010), Variable selection using Adaptive Nonlinear Interaction Structures in High dimensions (VANISH) (Radchenko and James, 2010), Strong Hierarchical Lasso (HIERNET) (Bien et al., 2013), Group Lasso Interaction Network (GLINTERNET) (Lim and Hastie, 2015), Framework for Modeling Interactions with a Convex Penalty (FAMILY) (Haris et al., 2016a), Group Regularized Estimation under Structural Hierarchy (GRESH) (She et al., 2018), and Sparse Additive Interaction Learning (SAIL) (Bhatnagar et al., 2018). The ideas are mainly to apply separate penalties (using the L^1, L^2, or L^∞ norm, depending on the method) to coefficients of main terms and interactions. However, these methods focus only on interaction selection and it is not trivial to incorporate selection rules where interactions are not included.

To incorporate a wider spectrum of selection rules, it is possible to extend the group Lasso, further relaxing the non-overlapped groups assumption. Related research started from a special structure of groups which required that a smaller group must be nested in a larger one. (Zhao et al., 2009; Jenatton et al., 2011b; Kim and Xing, 2012) Notably, Mairal et al. (2010) developed a method (which we call "overlapping group Lasso") that has no restriction on the structure of the groups. In Chapter 5, we extend this method to accommodate data containing survival outcomes and time-dependent covariates. We first introduce the method in this chapter below.

The overlapping group Lasso solves

$$\min_{\boldsymbol{\beta}} \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{\beta}\boldsymbol{X}\|_2^2 + \lambda \sum_{\mathbb{g} \in \mathbb{G}} \omega_{\mathbb{g}} \left\|\boldsymbol{\beta}_{|\mathbb{g}}\right\|_\infty .$$

Note that the penalty here is the sum of L^∞ norms (which is the maximum value of the vector $\boldsymbol{\beta}_{|\mathbb{g}}$) rather than the L^2 norm in group Lasso. Both norms can achieve the goal of forcing a group of variables to be selected collectively, but the L^∞ norm is piece-wise linear with respect to the vector inside of the norm.

Generalizing, we replace $\frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{\beta}\boldsymbol{X}\|_2^2$ with $\ell(\boldsymbol{\beta})$, defined as a generic continuously differentiable convex function. We then describe how to solve the above optimization problem. Because the penalty term is non-differentiable, Mairal et al. (2010) proposed to utilize the proximal method (Moreau, 1962) to overcome this issue. In each iteration of the proximal method, instead of updating the estimate with respect to the gradient of the objective function (as gradient descent methods do), it updates estimates that stay close to the gradient update for the differentiable function $\ell(\boldsymbol{\beta})$, while also making the non-differentiable penalty term small. (Beck and Teboulle, 2009; Nesterov, 2013a) It is proven to converge at fairly fast rates. (Nesterov, 2007; Beck and Teboulle, 2009; Beck, 2017) More specifically, by the Taylor expansion, in each iteration, denoting the value of $\boldsymbol{\beta}$ from the last update as $\tilde{\boldsymbol{\beta}}$, it solves

$$\min_{\boldsymbol{\beta}} \ell(\tilde{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^\top \nabla \ell(\tilde{\boldsymbol{\beta}}) + \lambda \sum_{\mathbb{g} \in \mathbb{G}} \omega_{\mathbb{g}} \left\|\boldsymbol{\beta}_{|\mathbb{g}}\right\|_\infty + \frac{L}{2} \left\|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\right\|_2^2,$$

where $L > 0$ is an upper bound on the Lipschitz constant of the gradient of $\ell, \nabla\ell$. The above problem can be further written as

$$\min_{\boldsymbol{\beta}} \frac{1}{2} \|\boldsymbol{u} - \boldsymbol{\beta}\|_2^2 + \frac{\lambda}{L} \sum_{\mathbb{g} \in \mathbb{G}} \omega_{\mathbb{g}} \left\|\boldsymbol{\beta}_{|\mathbb{g}}\right\|_\infty , \tag{2.7}$$

where $\boldsymbol{u} = \tilde{\boldsymbol{\beta}} - \frac{1}{L}\nabla\ell(\tilde{\boldsymbol{\beta}})$. Mairal et al. (2010) further proposed to solve its dual, which is

$$\min_{\boldsymbol{\xi}} \frac{1}{2}\left\|\boldsymbol{u} - \sum_{\mathbb{g}\in\mathbb{G}} \boldsymbol{\xi}_{|\mathbb{g}}\right\|_2^2, \text{ s.t. } \forall \mathbb{g} \in \mathbb{G}, \left\|\boldsymbol{\xi}_{|\mathbb{g}}\right\|_1 \leqslant \lambda\omega_{\mathbb{g}} \text{ and } \boldsymbol{\xi}_{|\mathbb{g},j} = 0 \text{ if } j \notin \mathbb{g}. \tag{2.8}$$

Denote the solution of (2.7) and (2.8) by $\boldsymbol{\beta}^*$ and $\boldsymbol{\xi}^*$ respectively. Then $\boldsymbol{\beta}^* = \boldsymbol{u} - \boldsymbol{\xi}^*$.

However, due to the nature of overlapping groups, solving (2.8) is still non-trivial. Mairal et al. (2010) further proposed to convert the problem to a quadratic min-cost flow problem (Hochbaum and Hong, 1995), which has been studied in operational research. One can use the network flow algorithm to solve the quadratic min-cost flow problem. Mairal et al. (2010) proposed a novel algorithm based on the network flow algorithm to solve (2.8), which was shown to have a better performance than the classic network flow algorithm in practice. It was proven that the proposed algorithm solves (2.8) and converges in a finite and polynomial number of operations. However, whether the estimator has the oracle property is currently unknown.

The network flow algorithm was developed and applied mainly in engineering and machine learning contexts (Varoquaux et al., 2011; Jenatton et al., 2012; Zhou et al., 2012; Mairal et al., 2014). Understanding the algorithm requires the knowledge of graph models (Gallo et al., 1989; Babenko and Goldberg, 2006) and thus remains challenging for researchers who do not have sufficient background in operational research and machine learning. Therefore, the overlapping group Lasso has not been broadly recognized and implemented in other fields such as biostatistics. In Chapter 5, as a component of the extension of the overlapping group Lasso, we present the algorithm in a more straightforward way to reach a broader audience, which avoids knowledge of graph models.

Depending on the grouping structure and the tuning parameters, the selected variables from the application of the overlapping group Lasso are the remaining variables that do not belong to the groups of variables whose coefficients are shrunk to zero. Therefore, the overlapping

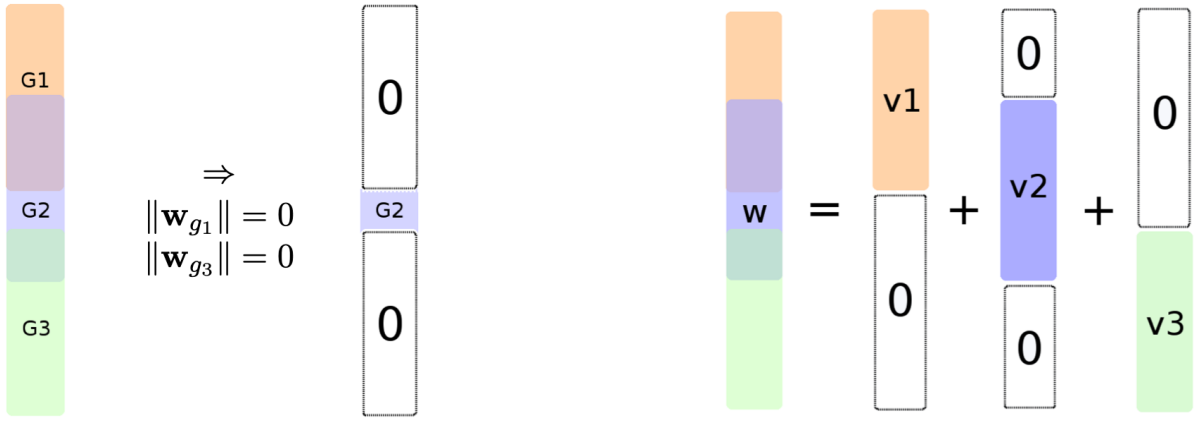
group Lasso can respect some of the selection rules in the if-then rules family (in Chapter 3, we use $\rightarrow$ to code the logic "if-then"). For example, "if at least one variable in a set is selected, then all the variables in another set must be selected". This corresponds to the strong heredity when the first set contains only one variable. However, how to define the grouping structure to respect such selection rules has not yet been well-studied.

The overlapping group Lasso is still not versatile enough to respect all types of selection rules. In addition, the overlapping group Lasso, just like Lasso and the group Lasso, may always arrive at results such as "no variable is selected" and "all variables are selected". In other words, the overlapping group Lasso cannot respect the selection rule which leads to excluding the above possible results.

2.2.9 Latent overlapping group Lasso

While the overlapping group Lasso selects variables by excluding the union of groups of variables whose coefficients are shrunk to zero, Obozinski et al. (2011a) developed a method (which we called "latent overlapping group Lasso") that can select variables by selecting the union of the groups of variables whose coefficients are not shrunk to zero. Figure 2.3 (Obozinski et al., 2011a) illustrates the two strategies of selection. It turns out that this strategy of selection can respect more types of selection rules. In Chapter 4, we use this method to show how to apply the framework established in Chapter 3 in identifying predictors of major bleeding among hospitalized hypertensive patients using oral anticoagulants for atrial fibrillation. The latent overlapping group Lasso solves

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{\beta}\boldsymbol{X}\|_2^2 + \lambda \sum_{\mathbb{g}\in\mathbb{G}} \omega_{\mathbb{g}} \left\|\boldsymbol{\alpha}_{|\mathbb{g}}\right\|_2 \quad \text{s.t.} \quad \sum_{\mathbb{g}\in\mathbb{G}} \boldsymbol{\alpha}_{|\mathbb{g}} = \boldsymbol{\beta},$$



(a) the overlapping group Lasso (b) the latent overlapping group Lasso

Figure 2.3: Two strategies of selection. Given groups G_1 (corresponds to $\mathbb{g}_1$), G_2, and G_3 in different colors, the coefficients are denoted by w_{g_1} (corresponds to $\boldsymbol{\beta}_{|\mathbb{g}_1}$), w_{g_2}, and w_{g_3}, and the latent variables are denoted by v1 (corresponds to $\boldsymbol{\alpha}_{|\mathbb{g}_1}$), v2, and v3. (a): the overlapping group Lasso, when the coefficients of variables in G_1 and G_3 are shrunk to zero, the selected variables are the left variables in G_2 but not in G_1 or G_3 (the variables in violet only). (b): the latent overlapping group Lasso, when v1 and v3 are non-zero, the selected variables are the union of the variables whose coefficients correspond to the positions of v1 and v3, the unselected variables are the variables in G_2 but not in G_1 or G_3. (Obozinski et al., 2011a)

where $\boldsymbol{\alpha}$ is a latent variable set, of size p. Equivalently, it can also be formulated as an optimization problem that minimizes $\boldsymbol{\beta}$ only

$$\min_{\boldsymbol{\beta}} \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{\beta}\boldsymbol{X}\|_2^2 + \lambda\Omega(\boldsymbol{\beta}), \quad \Omega(\boldsymbol{\beta}) \triangleq \min_{\boldsymbol{\alpha}, \sum_{\mathbb{g}\in\mathbb{G}} \boldsymbol{\alpha}_{|\mathbb{g}} = \boldsymbol{\beta}} \sum_{\mathbb{g}\in\mathbb{G}} \omega_{\mathbb{g}} \left\|\boldsymbol{\alpha}_{|\mathbb{g}}\right\|_2 .$$

The associated algorithm combines the block-coordinate decent (Meier et al., 2008) and the working set strategy (Roth and Fischer, 2008). In addition, the method can adapt SCAD and MCP penalties by replacing the β_j in (2.5) and (2.6) to $\left\|\boldsymbol{\alpha}_{|\mathbb{g}}\right\|_2$. The `R` package `grpregOverlap` can implement all three of these penalties. The latent overlapping group Lasso can achieve selection consistency when certain conditions are satisfied, but if it possess the oracle property is unknown. The choice of weight is more important in the latent overlapping group Lasso because it significantly affects the selection consistency and the false negative and false positive rates. Obozinski et al. (2011a) discuss in length how to set

the weights in several scenarios. We omit the details here.

Due to the nature of this selection strategy, the latent overlapping group Lasso can respect some selection rules that the overlapping group Lasso cannot, such as weak heredity. However, it shares the same limitation that all the above techniques have, that is, "no variable is selected" and "all variables are selected" are always possible results.

2.3 Survival analysis

2.3.1 General formulation of survival analysis

In many fields, especially in biological and epidemiological research, a time-to-event variable (denoted by T) is often of interest as the outcome. (Bartelink et al., 2001; Primrose et al., 2014; Andell et al., 2014; Banankhah et al., 2015) Analysis of such variables as the outcome is referred to as survival analysis. To characterize the distribution of time-to-event variables, we next introduce some functions.

Let $f(t)$ and $F(t)$ denote the probability density function and cumulative density function of T, respectively. The survival function

$$S(t) = 1 - F(T) = P(T > t) = \int_t^{\infty} f(x)dx$$

is defined as the probability that the event occurs after a time t (the probability of survival at t). (Kalbfleisch and Prentice, 2011) Another way to characterize the distribution of T is through the hazard function $h(t)$, which is the instantaneous risk of the event occuring at t conditional on event-free survival until t:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leqslant T < t + \Delta t | T > t)}{\Delta t} = \frac{f(t)}{S(t)}.$$

The cumulative hazard function, $H(t) = \int_0^t h(x)dx$, is the cumulative risk from time zero to

t. (Cleves et al., 2008)

Interestingly, the above functions have some one-to-one relationships (Kalbfleisch and Prentice, 2011)

$$h(t) = \frac{F'(t)}{S(t)} = -\frac{S'(t)}{S(t)} = \frac{f(t)}{S(t)} \quad \text{and} \quad S(t) = \exp\{-H(t)\}.$$

Censoring is an important component in survival analysis. It occurs when the time-to-event cannot be exactly recorded. Right censoring, the most common type of censoring, refers to the cases where the event is unobserved but is known to occur after a certain time point. Reasons for right censoring include loss to follow-up or termination of the study before the subject has the event (also known as administrative censoring). In the presence of right censoring, the time-to-event variable T is not observed for all subjects. Denote the observed time to either censoring or the event by U, and the censoring indicator by δ (binary, 1 if the time-to-event is observed, 0 otherwise). Then the outcome can be coded as a pair (U, δ). That is, if $\delta = 1$ then $T = U$ or if $\delta = 0$ then $T > U$.

For a dataset consisting of n subjects, if the event occurs for subject i, its contribution to the likelihood is the density at the time-to-event. Otherwise, assuming non-informative censoring (Kalbfleisch and Prentice, 2011), the contribution of the censored subject would be the survival probability, as we only know the subject survives until the censoring time. The likelihood function can then be written as

$$\prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} = \prod_{i=1}^{n} h(t_i)^{\delta_i} S(t_i).$$

Many methods can model time-to-event outcomes under different assumptions. Of these, the Cox proportional hazards model (Cox, 1972) is the most popular and has become the standard method in medical research. It is a semi-parametric model which does not require assuming the distribution of the event times. Next, we first introduce the Cox model and then

review some penalization methods that can be applied to the estimation of the parameters of the model. We then illustrate how time-dependent covariates can be incorporated in the Cox model.

2.3.2 Cox proportional hazards model

The Cox proportional hazards model is specified as

$$h(t|\boldsymbol{X}) = h_0(t)\exp(\boldsymbol{X}\boldsymbol{\beta}),$$

where $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ and $\boldsymbol{\beta} \in \mathbb{R}^p$ are the covariate matrix and the vector of coefficients, respectively. The baseline hazard function $h_0(t)$ is a function of t that describes the hazard at t when $\boldsymbol{X} = \mathbf{0}$. Essentially, this model assumes that 1) the hazard of a subject at t is proportional to the baseline hazard and 2) the logarithm of the hazard can be written as a linear combination of the covariates. The coefficient of one covariate x can then be interpreted as the logarithm of the hazard ratio for a one unit increase in x, holding other covariates constant.

Cox (1972) proposed to use the partial likelihood for estimation. Let $j = 1, \ldots, m$ be the index of ordered unique observed follow-up times in the dataset, and $t_1 \leqslant t_2 \leqslant \ldots \leqslant t_m$ be an ordered list of unique time-to-events. There are d_j tied events occurring at the jth distinct survival time. Let D_j and R_j be the index sets of subjects whose event occurred at time t_j or are at risk at time t_j, respectively. Being "at risk" at time t means that the event has not yet happened before time t, and the subject was not censored before or at time t. The partial likelihood can then be written as

$$L(\boldsymbol{\beta}) \approx \prod_{j=1}^{n} \left\{ \frac{\exp(\boldsymbol{X}_i\boldsymbol{\beta})}{\sum_{l\in R_i}\exp(\boldsymbol{X}_l\boldsymbol{\beta})} \right\}^{\delta_i}.$$

Using the Breslow approximation to accommodate the tied events (Breslow, 1974), the partial likelihood and the log partial likelihood can be approximated as

$$L(\boldsymbol{\beta}) \approx \prod_{j=1}^{m} \frac{\exp\left\{\left(\sum_{l\in D_j} \boldsymbol{X}_l\right)\boldsymbol{\beta}\right\}}{\left\{\sum_{l\in R_j}\exp(\boldsymbol{X}_l\boldsymbol{\beta})\right\}^{d_j}}, \quad \text{and}$$

$$\log\{L(\boldsymbol{\beta})\} \approx \sum_{j=1}^{m}\left[\left(\sum_{l\in D_j}\boldsymbol{X}_l\right)\boldsymbol{\beta} - d_j \log\left\{\sum_{l\in R_j}\exp(\boldsymbol{X}_l\boldsymbol{\beta})\right\}\right],$$

respectively. It is well known that the log partial likelihood is a concave function. (Elston et al., 2002) Define the partial likelihood score as the derivative of the log partial likelihood with respect to $\boldsymbol{\beta}$. The estimated $\boldsymbol{\beta}$ can then be obtained by setting the partial likelihood score to 0.

While most of the variable selection techniques mentioned in section 2.2 are developed and applied in the context of multiple linear regression, many of them (except for the best subset selection via optimization, the exclusive Lasso, and the overlapping group Lasso) have been extended to the Cox model and the theoretical properties were investigated. (Tibshirani, 1997; Zhang and Lu, 2007; Ma et al., 2007; Wang et al., 2009; Benner et al., 2010; Breheny and Huang, 2011; Bradic et al., 2011; Sun et al., 2014; Tang et al., 2019) When the covariates are time-dependent, Wallace (2014) proposed a time-dependent tree-structured model for variable selection with a survival outcome, but no selection rule can be incorporated.

2.3.3 Time-dependent Cox model

In many research domains, data are collected longitudinally, which can provide non-negligible information for predicting the outcomes. (Fisher and Lin, 1999) For instance, in medical research, treatment assignment, blood pressure, weight, disease history, and hospitalization information may be recorded at selected times over the course of a patient's follow-up. The time-varying values of these characteristics over time may be helpful in predicting the time-

to-event. Next, we introduce how to accommodate time-dependent covariates in the Cox model.

For ease of notation, we code all covariates as time-dependent, i.e. $\boldsymbol{X}_i(t) = (x_{i1}(t), x_{i2}(t), \cdots, x_{ip}(t))$, $i = 1, \ldots, n$, though some may be fixed at baseline or constant over time. We assume that the hazard of subject i at time t only depends on the covariates through their current values at time t. Then the time-dependent Cox model (Christensen et al., 1986; Fisher and Lin, 1999) assumes that the conditional hazard function is $h_i\{t|\boldsymbol{X}_i(t)\} = h_0(t)\exp\{\boldsymbol{X}_i(t)\boldsymbol{\beta}\}$. The partial likelihood and the log partial likelihood can be approximated as

$$L(\boldsymbol{\beta}) \approx \prod_{j=1}^{m} \frac{\exp\left[\{\sum_{l\in D_j} \boldsymbol{X}_l(t_j)\}\boldsymbol{\beta}\right]}{\left[\sum_{l\in R_j}\exp\{\boldsymbol{X}_l(t_j)\boldsymbol{\beta}\}\right]^{d_j}},$$

and

$$\log\{L(\boldsymbol{\beta})\} \approx \sum_{j=1}^{m}\left(\left\{\sum_{l\in D_j}\boldsymbol{X}_l(t_j)\right\}\boldsymbol{\beta} - d_j\log\left[\sum_{l\in R_j}\exp\{\boldsymbol{X}_l(t_j)\boldsymbol{\beta}\}\right]\right), \tag{2.9}$$

respectively. Estimating the coefficients of time-dependent covariates is similar to estimation under the standard Cox model.

While other specifications of time-dependent Cox models exist (de Bruijne et al., 2001), in Chapter 5, we extend the overlapping group Lasso to this form of time-dependent Cox model.

Chapter 3

A general framework for identification of permissible variable subsets and development of structured variable selection methods

Preamble to Manuscript 1. As explained in Chapter 2.1.2, inherent covariate structures are often present in data analysis. Much research has been done to accommodate such structures. However, no theory or method addresses the problem in full generality. In this chapter, we provide a new perspective on structured variable selection that can incorporate universal structural constraints. We first develop a mathematical language for constructing selection dependencies according to the corresponding covariate structures, which we call "selection rules". Then we derive an algorithm to determine all permissible subsets of covariates that respect the selection rules, which we call "selection dictionary". We then show that the theoretical framework can help to 1) identify the grouping structure in existing penalized regression methods, and 2) formulate structured variable selection into Mixed-Integer

Optimization (MIO) problems which can be solved by existing software. The significance of the theoretical framework and its two applications are then discussed. This manuscript was submitted to *Annals of Statistics*.

Note that the supplementary material for this chapter can be found in Appendix A.

A general framework for identification of permissible variable subsets and development of structured variable selection methods

Guanbo Wang[1], Mireille E. Schnitzer[2,3,1], Tom Chen[4], Rui Wang[4,5], and Robert W. Platt[6,1]

[1]*Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, QC, Canada*

[2]*Faculté de pharmacie, Université de Montréal, Montréal, QC, Canada*

[3]*Département de médecine sociale et préventive, ESPUM, Université de Montréal, Montréal, QC, Canada*

[4]*Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, MA, USA*

[5]*Harvard T. H. Chan School of Public Health, Harvard University, Boston, MA, USA*

[6]*Department of Pediatrics, McGill University, Montréal, QC, Canada*

This chapter contains the corresponding paper to be submitted to *Annals of Statistics*

Abstract

In variable selection, a selection rule that prescribes the permissible sets of selected variables (called a "selection dictionary") is desirable due to the inherent structural constraints among the candidate variables. The methods that can incorporate such restrictions can improve model interpretability and prediction accuracy. Penalized regression can integrate selection rules by assigning the coefficients to different groups and then applying penalties to the groups. However, no general framework has been proposed to formalize selection rules and their applications. In this work, we establish a framework for structured variable selection that can incorporate universal structural constraints. We develop a mathematical language for constructing arbitrary selection rules, where the selection dictionary is formally defined. We show that all selection rules can be represented as a combination of operations on constructs, which can be used to identify the related selection dictionary. One may then apply some criteria to select the best model. We show that the theoretical framework can help to identify the grouping structure in existing penalized regression methods. In addition, we formulate structured variable selection into MIO problems which can be solved by existing software. Finally, we discuss the significance of the framework in the context of statistics.

3.1 Introduction

Variable selection has become an important technique in statistics and data science, especially with large-scale and high-dimensional data becoming increasingly available. Variable selection can be used to identify covariates that are associated with or predictive of the outcome, remove spurious covariates, and improve prediction accuracy. (Guyon and Elisseeff, 2003; Reunanen, 2003; Wasserman and Roeder, 2009) General techniques to conduct statistical variable selection include best subset selection, penalized regression, and nonparametric approaches like random forest. (Heinze et al., 2018; Chowdhury and Turin, 2020)

When selecting variables for the purpose of developing an interpretable model, understanding and formalizing the structure of covariates can lead to more interpretable variable selection. Covariates may have a structure due to

1. Variable type. For example, when including a categorical variable in a regression model, each non-reference category is represented by a binary indicator. It may be desirable to collectively include or exclude these binary indicators as a group.

2. Variable hierarchy. For example, one may define a hierarchical structure for sets of covariates $\mathbb{A}$ and $\mathbb{B}$ such that if $\mathbb{A}$ is selected, then $\mathbb{B}$ must be selected. One application is interaction selection with strong heredity (Haris et al., 2016a), that is "the selection of an interaction term requires the inclusion of all main effect terms". A second application is when one covariate is a descriptor of another, such as medication dose (0-10mL) and medication usage (yes/no).

Such restrictions on the resulting model, which we call "selection rules", can be incorporated in the variable selection process so that the resulting model satisfies the rules after statistical selection is carried out. Practitioners can define any selection rule based on their *a priori* knowledge of the covariate structure.

Lasso (Tibshirani, 1996) and best subset selection via optimization (Bertsimas et al., 2016)

are two approaches that do not restrict the composition of the resulting model. However, certain types of selection dependencies have been integrated into covariate coefficient penalization. For instance, group Lasso (Yuan and Lin, 2006) can select a group of variables collectively. Exclusive Lasso (Campbell and Allen, 2017) can achieve within group selection by selecting at least one variable in each group. Overlapping group Lasso (Mairal et al., 2010) and latent overlapping group Lasso (Obozinski et al., 2011a) achieve hierarchical selection by requiring that a group of variables be selected when another group has been selected. Note that each method can respect one or more types of selection rules.

In the absence of a selection rule, there are no restrictions on the potential resulting model – all combinations of covariates are allowed to be selected. Given a selection rule, this set becomes restricted to only include combinations of covariates that respect the selection rule. We call the set of all potential sets of covariates a "selection dictionary". Currently, there is no unifying framework for structured selection or generic algorithm that can generate the selection dictionary respecting any given selection rule. Manually constructing selection rules for specific settings can be tedious and prone to errors, even in low-dimensional covariate settings and especially when the selection rules are complicated. In this work, we develop a general framework for variable selection that can represent any selection rule in a formal mathematical language, allowing us to catalogue the selection dictionary for any given selection rule. In low-dimensional settings, once the selection dictionary is known, goodness-of-fit measures like AIC or BIC, or prediction accuracy measures like cross-validated prediction error, can be used to select the best model out of all allowable models.

Penalized regression offers an alternative to variable selection and is especially useful for higher covariate dimensions. Given a penalized regression method, an existing gap in the literature is a general approach to grouping covariates that enforces respect of a selection rule. While grouping structures have been developed for specific rules (Lim and Hastie, 2015; Yuan and Lin, 2006; Mairal et al., 2010; Jenatton et al., 2011a; Campbell and Allen,

2017; Yan et al., 2017), how to define grouping structures for more complex selection rules has not been well-studied. For example, suppose that we have three categorical variables and are also interested in two-way interactions between these variables. Our selection rule is to select an interaction only if at least one of main terms is also selected. We also want to select the dummy variables as a group, both as main terms and in the interactions. How to group the three main terms and their interactions in latent overlapping group Lasso is not trivial. In this work, we identify the sufficient and necessary condition of a grouping structure that, when coupled with latent overlapping group Lasso, will respect any selection rule. This result can then guide us in how to group variables.

In addition, current penalized regression methods for structured variable selection cannot respect all possible selection rules, in particular those that control the number of selected covariates. Based on the defined framework, we formulate solvable optimization problems based on penalization by the L^0 norm which open the door for the development of methods that can respect broader classes of selection rules.

This paper is organized as follows. In section 3.2, we give an overview of the results presented in the rest of the paper. In section 3.3 we formally introduce the language for constructing selection rules, and describe how to express an arbitrary selection rule and how to determine the selection dictionary. Next, in section 3.4, we apply the framework to existing penalized regression methods by defining penalization structure and giving the sufficient and necessary condition for respecting a rule with latent overlapping group Lasso. In section 3.5, we propose selection rule-based variable selection techniques that can respect the generic unit rules and any binary operation on these. Lastly, we discuss the significance of the proposed framework.

3.2 Overview

In this section, we outline and explain our framework, and its applications in grouping structure identification and constructing selection rule-based variable selection methods. More rigorous development along with additional examples are given in sections 3.3, 3.4 and 3.5.

Suppose that we have a set of candidate variables $\mathbb{V}$. We define a selection rule on this set as the selection dependencies among all candidate variables. For example, consider a study where we want to investigate which of the following variables should be included in a model for blood pressure: age (A), age squared (A^2), and race as a categorical variable with 3 levels, represented by dummy variables B_1 and B_2. We are also interested in the interaction of age with race (AB_1, AB_2). So we have $\mathbb{V} = \{A, A^2, B_1, B_2, AB_1, AB_2\}$. In this example, standard statistical practice requires that the resulting model must satisfy a selection rule defined by the following three conditions: 1) if the interaction is selected, then both age and race must be selected, 2) if age squared is selected, then age must be selected, 3) the dummy variables representing race must be collectively selected, and 4) the two categorical interaction terms must also be collectively selected. The combination of these four rules is the selection rule that must be respected.

We next define a selection dictionary as the set that contains all subsets of $\mathbb{V}$ that respect the selection rule. When we say a dictionary respects a selection rule, we mean the dictionary is congruent to the selection rule in the sense that the selection dictionary contains all (rather than some) subsets of $\mathbb{V}$ that respect the selection rule. Theorem 1 states that every selection rule has a unique dictionary. The dictionary for the above example would be $\{\emptyset\ ,\ \{A\}\ ,\ \{B_1, B_2\}\ ,\ \{A, B_1, B_2\}\ ,\ \{A, A^2\}\ ,\ \{A, A^2, B_1, B_2\}\ ,\ \{A, B_1, B_2, AB_1, AB_2\}\ ,\ \{A, A^2, B_1, B_2, AB_1, AB_2\}\}$. Despite a total of 64 possible subsets of $\mathbb{V}$, there are only 8 possible models that can be selected under this rule.

We are interested in the general problem of finding the selection dictionary given an arbitrary

selection rule. We start by defining unit rules as the building blocks of selection rules. For a given set of candidate variables $\mathbb{V}$ with $\mathbb{F} \subseteq \mathbb{V}$, a unit rule is a selection rule of the form "select a number of variables in $\mathbb{F}$." The unit rule depends on the set $\mathbb{C}$ which contains the numbers of variables that are allowed to be selected from $\mathbb{F}$. In our running example, one unit rule is "select zero or two variables from the set $\mathbb{F} = \{B_1, B_2\}$". This is equivalent to saying that B_1 and B_2 must be selected together, i.e. select neither or both.

In Theorem 2, we give a formula for the dictionary related to any given unit rule. This formula shows that the dictionary is all unique unions of sets 1) of variables in $\mathbb{F}$ where the number of variables is in $\mathbb{C}$ and 2) of variables outside of $\mathbb{F}$. Applying this formula, we can see that the unit rule "select zero or two variables (that is, $\mathbb{C} = \{0, 2\}$) from the set $\mathbb{F} = \{B_1, B_2\}$" has a dictionary that is the set incorporating $\emptyset$ and $\{B_1, B_2\}$ and all unions of $\emptyset$ and $\{B_1, B_2\}$ with any of the other elements in $\mathbb{V}$, respectively.

We then define five useful operations on selection rules in Table 3.2. For example, $\wedge$ being applied to two selection rules indicates that both of the selection rules must be respected. An arrow $\rightarrow$ indicates if the selection rule on the left hand side is being respected, then the selection rules on the right hand side must be respected. For each operation, we can show how the operation on selection rules is related to an operation of the respective dictionaries. Therefore if we are combining or constructing more complex rules from operations on unit rules, we can always derive the resulting dictionary. Our most important result is Theorem 3, stating that through operations on unit rules, we can derive any rule.

To illustrate these ideas in our running example, define unit rules 1) $\mathfrak{u}_1$: "select zero or two variables in $\{AB_1, AB_2\}$," 2) $\mathfrak{u}_2$: "select zero or two variables in $\{B_1, B_2\}$," 3) $\mathfrak{u}_3$: "select two variables in $\{AB_1, AB_2\}$," 4) $\mathfrak{u}_4$: "select three variables in $\{A, B_1, B_2\}$," 5) $\mathfrak{u}_5$: "select one variable in $\{A^2\}$," 6) $\mathfrak{u}_6$: "select one variable in $\{A\}$". The same selection rule that we defined when we introduced the example can be expressed through operations on these unit rules as: $(\mathfrak{u}_1 \wedge \mathfrak{u}_2) \wedge (\mathfrak{u}_3 \rightarrow \mathfrak{u}_4) \wedge (\mathfrak{u}_5 \rightarrow \mathfrak{u}_6)$.

Section 3.4 involves describing how to assign group-specific penalties in penalized regression in order to respect a selection rule with the aid of the determined selection dictionary. By the properties of these penalized regression methods, a given method with a given grouping structure (i.e. grouping of variables in the penalty terms) results in a unique dictionary. This dictionary contains all subsets of $\mathbb{V}$ that could potentially result from the application of the penalization method on the dataset. So if the dictionary related to a penalized regression method with a specific grouping structure is equal to a selection rule's dictionary, then the method is "congruent" to the rule. Essentially, the grouping structure in the penalty terms in a specific penalization regression determines the selection rule that the method can respect.

Focusing on latent overlapping group Lasso, Theorem 4 gives the sufficient and necessary condition under which a grouping structure respects a selection rule. This lets us find a grouping structure that respects the selection rule. In the example, one penalization structure that is congruent with the selection rule is latent overlapping group Lasso paired with the grouping structure $\{\{A\}, \{A, A^2\}, \{B_1, B_2\}, \{A, B_1, B_2, AB_1, AB_2\}\}$.

Existing penalized regression methods were developed for respecting a certain type(s) of selection rules. In Section 3.5, inspired by the method that constructing the selection rules and mixed-integer optimization (MIO), we developed optimization problems that can directly control the number of variables being selected in a subset of candidate variables, so that it can respect unit rules and also their operations. The proposed problems can be solved by existing software, and thus enriches the scope of a priori knowledge that can be incorporated in variable selection.

Section 4.5 explains the significance of the framework within the broader scope of statistical variable selection methods. As an example, we describe the limitation of (latent) overlapping group Lasso in respecting selection rules.

3.3 Selection rules and selection dictionaries

In this section, we introduce the mathematical language for expressing selection rules, which enables us to design algorithms to incorporate selection dependencies into model selection. Two fundamental concepts are being introduced first: the selection rule and its dictionary. Then we introduce unit rules and operations on unit rules as the building blocks of selection rules. We show that we can construct any selection rule from unit rules and also derive the unit dictionary from set operations on the dictionaries belonging to the unit rules. Finally, we investigate some properties of the resulting abstract structures.

Unless specified otherwise, we use normal math (for example F), blackboard bold ($\mathbb{F}/\mathbb{f}$), Fraktur lowercase ($\mathfrak{f}$), and calligraphy uppercase fonts ($\mathcal{F}$) to represent a random variable, set, rule, and operator respectively. $\mathcal{P}(\mathbb{F})$ represents the power set (collection of all possible subsets) of $\mathbb{F}$, $\mathcal{P}^2(\mathbb{F})$ denotes the power set of the power set of $\mathbb{F}$, and $|\mathbb{F}|$ represents the cardinality of $\mathbb{F}$. The maximum integer of a set of integers $\mathbb{F}$ is denoted by $\max(\mathbb{F})$. We say two sets are equivalent if they contain the same elements, regardless their multiplicity. For example $\{A, A, B, B, C, C\} = \{A, B, C\}$. Graphs are helpful to show the dependencies among candidate covariates. For example, an arrow in a graph can indicate that the children nodes are constructed based on their parent nodes.

We take two examples to illustrate the concepts throughout this section.

Example 1. Suppose that we have 4 candidate variables, $\mathbb{V} = \{A, B, C, D\}$, that have no structural relationship. The corresponding graph is shown in Figure 3.1.

Example 2. Suppose we have three variables: a continuous variable A and a three-level categorical variable B (represented by two dummy indicators B_1 and B_2). We also consider their interactions represented by AB_1 and AB_2. The corresponding graph is shown in Figure 3.2 with nodes $\mathbb{V} = \{A, B_1, B_2, AB_1, AB_2\}$. The arrows indicate that the child nodes are derived from their parents.

$\mathfrak{r}_i$	**Selection dependencies**	$\mathbb{D}_{\mathfrak{r}_i}$ **Selection dictionaries**
$\mathfrak{r}_1$	Select at least one variable in $\{A, B\}$.	$\big\{\{A\}, \{B\}, \{A, B\}, \{A, C\}, \{B, C\}, \{A, B, C\}\big\}$
$\mathfrak{r}_2$	If A is selected, then B must be selected.	$\big\{\emptyset, \{B\}, \{A, B\}, \{C\}, \{B, C\}, \{A, B, C\}\big\}$
$\mathfrak{r}_3$	Respect both $\mathfrak{r}_1$ and $\mathfrak{r}_2$.	$\big\{\{B\}, \{A, B\}, \{B, C\}, \{A, B, C\}\big\}$

Table 3.1: Examples of selection rules and their dictionaries, $\mathbb{V} = \{A, B, C\}$

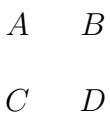


Figure 3.1: Graph for Example 1

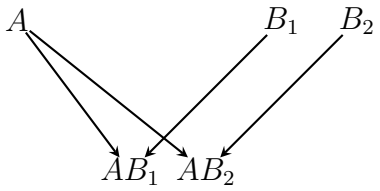


Figure 3.2: Graph for Example 2

Next, we introduce the concept of selection rule.

Definition 1 (Selection rule). *A* ***selection rule*** $\mathfrak{r}$ *of* $\mathbb{V}$ *is defined as selection dependencies among the variables in* $\mathbb{V}$.

The selection dependency is a general concept regarding limitations on which combinations of variables are allowed to be selected into a model.

Table 3.1 gives some examples of selection rules for a covariate set $\mathbb{V} = \{A, B, C\}$. There may be many possible subsets of variables that respect a given selection rule. We define the set of all possible subsets respecting a given selection rule as the corresponding selection dictionary. Before introducing selection dictionary, we define a general dictionary first.

Definition 2 (Dictionary). *Given a finite set of candidate variables* $\mathbb{V}$*, a* ***dictionary*** $\mathbb{D} \subseteq \mathcal{P}(\mathbb{V})$ *of* $\mathbb{V}$ *is a set of subset(s) of* $\mathbb{V}$.

For example, a dictionary of candidate variables $\mathbb{V} = \{A, B, C, D\}$ can be $\{\{A\}, \{B\}\}$,

$\{\emptyset\ ,\ \{A, B, C, D\}\ ,\ \{A\}\}$ or $\mathcal{P}(\mathbb{V})$ etc.

DEFINITION 3 (Selection dictionary). *For a given* $\mathbb{V}$, ***selection dictionary*** $\mathbb{D}_{\mathfrak{r}}$ *is a dictionary that contains all subsets of* $\mathbb{V}$ *that respect the selection rule* $\mathfrak{r}$.

The definition stresses that all sets in a selection dictionary must be a subset of $\mathcal{P}(\mathbb{V})$. This allows us to list all "allowable" sets of variables that could result from a variable selection process respecting the selection rule. Some examples of selection dictionaries corresponding to selection rules are shown in Table 3.1. It is also possible to have selection rules that are *contradictory/incoherent*, meaning that they require more variables to be selected than the number of variables in the given set. In this case, we define the selection dictionary to be the $\emptyset$. For instance, if the selection rule is "select 3 variables from $\{A, B\}$," with $\mathbb{V} = \{A, B\}$, then the corresponding selection dictionary is $\emptyset$. When a selection rule is for example, "select 0 variables from $\{A, B\}$," which is coherent but trivial, then the selection dictionary is the empty set $\{\emptyset\}$.

By the definitions above, there is a mapping from a selection rule to a selection dictionary. The theorem below gives the uniqueness of the mapping with proof in Appendix A.1.

THEOREM 1. *Given a selection rule on a set, there is a unique selection dictionary.*

We say the unique selection dictionary is *congruent* to its selection rule, which is denoted by $\mathbb{D}_{\mathfrak{r}} \cong \mathfrak{r}$, or equivalently, $\mathfrak{r} \cong \mathbb{D}_{\mathfrak{r}}$. In our context, it is equivalently saying a selection dictionary respects a selection rule. However, it is possible that more than one selection rule results in the same selection dictionary. Therefore, we define an equivalence class of selection rules below.

DEFINITION 4 (Equivalence class of selection rules). *For a given candidate set* $\mathbb{V}$, *and given*

a selection rule $\mathfrak{r}_1$ *with selection dictionary* $\mathbb{D}$*, the* ***equivalence class*** *of* $\mathfrak{r}_1$*, denoted by* $\mathbb{R} \coloneqq \{\mathfrak{r} : \mathfrak{r} \cong \mathbb{D}\}$*, is a set of all selection rules in* $\mathbb{V}$ *that are congruent to the same selection dictionary.*

COROLLARY 1. *By Definition 4 and Theorem 1, there is a one-to-one mapping from an equivalence class of a selection rule to a selection dictionary.*

For a given (finite) $\mathbb{V}$, we define $\mathfrak{R}$ as the **selection rule space** containing all equivalence classes of selection rules on $\mathbb{V}$. Because the number of possible combinations of selected variables is finite, the number of possible dictionaries is finite. Thus, because of the one-to-one correspondence between dictionaries and rules, the space of rules $\mathfrak{R}$ is also finite.

The above definitions provide us with a broad view of the language for expressing selection rules generally. Next we introduce the grammar of this language which allows for the exploration of theoretical properties of selection rules and of further algorithmic development. We start by defining unit rules and their dictionaries, and then introduce the operations between unit rules. Then more complex selection rules can be assembled by unit rules and their operations, and the related selection dictionaries can be determined.

DEFINITION 5 (Unit rule and its dictionary). *For a given* $\mathbb{V}$ *and a given* $\mathbb{F} \subseteq \mathbb{V}$*, a* ***unit rule*** $\mathfrak{u}_{\mathbb{C}}(\mathbb{F})$ *is a selection rule, where the selection dependency takes the form "there are a number of variables in* $\mathbb{F}$ *to be selected", and the number of variables to be selected is constrained by* $\mathbb{C}$*, a set of numbers. A* ***unit dictionary*** $\mathbb{D}_{\mathfrak{u}}$ *is a dictionary that contains all subsets of* $\mathbb{V}$ *that respect the unit rule* $\mathfrak{u}_{\mathbb{C}}(\mathbb{F})$*.*

REMARK 1. *The set* $\mathbb{C}$ *is a set of numbers which constrains the number of variables to be selected in* $\mathbb{F}$*. For example, if* $|\mathbb{F}| = 3$ *and* $\mathbb{C}$ *is* $\{1\}$ *or* $\{0, 2\}$*, then the rule in* $\mathfrak{u}_{\mathbb{C}}(\mathbb{F})$ *translates to "there is one variable to be selected" or "there are zero or two variables to be selected" from*

$\mathbb{F}$, *respectively. Any* $\mathbb{C}$ *with elements greater than* $|\mathbb{F}|$ *would result in an incoherent unit rule because variable selection is done without replacement and so we cannot select more than the cardinality of the set.*

In the context where we investigate more than one unit rule for a given $\mathbb{V}$, we use $\mathfrak{u}_i$ to represent $\mathfrak{u}_{\mathbb{C}_i}(\mathbb{F}_i), i \geqslant 1$. Among the examples in Table 3.1, only $\mathfrak{r}_1$ is a valid unit rule, which can be expressed as $\mathfrak{u}_{\{1,2\}}(\{A, B\})$, and the unit dictionary $\mathbb{D}_{\mathfrak{u}} = \mathbb{D}_{\mathfrak{r}_1}$ is given.

Given a unit rule and its dictionary, there is a one-to-one mapping from the $\mathbb{F}$ in the unit rule to the corresponding unit dictionary, which is characterized by a unit function $f_{\mathbb{C}}$.

DEFINITION 6 (Unit function). *Each unit rule relates to a* ***unit function*** $f_{\mathbb{C}}$ *with some input* $\mathbb{F} \subseteq \mathbb{V}$. *A unit function* $f_{\mathbb{C}}$ *maps the subset* $\mathbb{F}$ *to the unit dictionary* $\mathbb{D}_{\mathfrak{u}} \in \mathcal{P}^2(\mathbb{V})$ *that respects* $\mathfrak{u}_{\mathbb{C}}(\mathbb{F})$.

Therefore, for a given unit rule, we can write $\mathfrak{u}_{\mathbb{C}}(\mathbb{F}) \cong \mathbb{D}_{\mathfrak{u}} = f_{\mathfrak{c}}(\mathbb{F})$. This is a valid function because a unit dictionary is defined as the set of all possible subsets of $\mathbb{V}$ that respect the unit rule; thus, given a fixed constraint and set $\mathbb{F}$, there is a unique dictionary output.

The following theorem characterizes unit functions, providing a formula for the unit dictionary, so that when a unit rule is given on $\mathbb{F}$, the corresponding unit dictionary can be determined.

THEOREM 2. *Each unit function in* $\mathbb{V}$ *is a* $\mathbb{C}$*-specific function* $f_{\mathbb{C}}(\cdot)$, *with domain* $\mathcal{P}(\mathbb{V})$, *defined by*

$$\mathbb{F} \mapsto \mathbb{D}_{\mathfrak{u}} = \begin{cases} \{\mathfrak{a} \cup \mathfrak{b} : \forall \mathfrak{a} \subseteq \mathbb{F} \ s.t. \ |\mathfrak{a}| \in \mathbb{C}, \forall \mathfrak{b} \subseteq \mathbb{V} \setminus \mathbb{F}\}, \ if \ |\mathbb{F}| \geqslant \max(\mathbb{C}) \\ \emptyset, \ otherwise, \end{cases}$$

where $\mathbb{F} \in \mathcal{P}(\mathbb{V})$.

When $|\mathbb{F}| \geqslant \max(\mathbb{C})$*, the unit rule* $\mathfrak{u}_{\mathbb{C}}(\mathbb{F})$ *is coherent, and the unit function is a bijection with domain* $\mathbb{M} = \{\mathbb{F},\ s.t.\ \mathbb{F} \in \mathcal{P}(\mathbb{V}), |\mathbb{F}| \geqslant \max(\mathbb{C})\}$ *and image* $\{\mathfrak{a} \cup \mathfrak{b}, \forall \mathfrak{a} \subseteq \mathbb{F}\ s.t.\ |\mathfrak{a}| \in \mathbb{C}, \forall \mathfrak{b} \subseteq \mathbb{V} \setminus \mathbb{F}, \forall \mathbb{F} \in \mathbb{M}\}$.

This means that when the unit rule is coherent, the unit dictionary contains all sets that are unions between a subset of $\mathbb{F}$ that respects the constraint $\mathbb{C}$ and a subset of the remaining covariates in $\mathbb{V}$ (excluding $\mathbb{F}$). The proof is given in Appendix A.2. Corollary 2 gives a special case of Theorem 2, characterizing the mapping of $\mathbb{V}$ to a unit dictionary by a unit function. Corollary 3 gives an interesting property of a unit dictionary. The proofs are direct consequences of Theorem 2.

COROLLARY 2. *When the input of a unit function is* $\mathbb{V}$*, with a constraint* $\mathbb{C}$ *resulting in a coherent unit rule, the resulting unit dictionary is* $f_{\mathbb{C}}(\mathbb{V}) = \{\mathfrak{n} \in \mathcal{P}(\mathbb{V}) : |\mathfrak{n}| \in \mathbb{C}\}$.

COROLLARY 3. *When* $\mathbb{C} \neq \{0\}$ *for a given coherent unit rule, the corresponding unit dictionary* $\mathbb{D}_{\mathfrak{u}}$ *satisfies* $\cup_i \mathbb{D}_{\mathfrak{u},i} = \mathbb{V}$*, where* $\mathbb{D}_{\mathfrak{u},i}$ *is the* i*th set in* $\mathbb{D}_{\mathfrak{u}}$.

To further investigate the relationships among unit functions with different constraints, we provide the following corollaries.

COROLLARY 4. *For a given* $\mathbb{F} \in \mathcal{P}(\mathbb{V})$*,* $f_{(\cdot)}(\mathbb{F})$ *is injective with respect to the argument* $\mathbb{C}$ *when at least one constraint (*$\mathbb{C}_1$ *or* $\mathbb{C}_2$*) results in a coherent rule applied to* $\mathbb{F}$*. That is,* $f_{\mathbb{C}_1}(\mathbb{F}) \neq f_{\mathbb{C}_2}(\mathbb{F})$ *whenever* $\mathbb{C}_1 \neq \mathbb{C}_2$*. This means that two distinct unit functions (related to two distinct unit rules) will result in different dictionaries even when the inputs are the same.*

COROLLARY 5. *For* $\mathbb{C} = \{0, \ldots, |\mathbb{F}|\}$ *then* $f_{\mathbb{C}}(\mathbb{F}) = \mathcal{P}(\mathbb{V}), \forall \mathbb{F} \subseteq \mathbb{V}$*. This means that when*

there is effectively no constraint on the selection (i.e. any number of variables can be selected), the unit dictionary is the power set of $\mathbb{V}$. *A consequence is that two different unit functions with nonrestrictive constraints can result in the same dictionary even when the inputs are different.*

The proofs of corollaries 4 and 5 are in Appendix A.3 and A.4, respectively.

The goal is to build selection rules out of unit rules. This will allow for an algorithm to determine the resulting selection dependencies and dictionary. To do this, we define some operations among selection rules. Because a unit rule is also a selection rule, the operations can be applied to unit rules.

DEFINITION 7 (Operations on selection rules). *Given selection rules on* $\mathbb{V}$, *define an* ***operation on selection rules*** $\mathcal{O}$ *as a function that maps a single selection rule or pair of selection rules to another selection rule* $\mathfrak{r}_{\mathcal{O}}$.

Operation	**Interpretation**	$\mathbb{D}_{\mathcal{O}}$
$\neg\mathfrak{r}_1$	$\mathfrak{r}_1$ is not being respected	$\mathcal{P}(\mathbb{V}) \setminus \mathbb{D}_{\mathfrak{r}_1}$
$\mathfrak{r}_1 \wedge \mathfrak{r}_2$	both $\mathfrak{r}_1$ and $\mathfrak{r}_2$ are being respected	$\mathbb{D}_{\mathfrak{r}_1} \cap \mathbb{D}_{\mathfrak{r}_2}$
$\mathfrak{r}_1 \vee \mathfrak{r}_2$	either $\mathfrak{r}_1$ or $\mathfrak{r}_2$, or both is/are being respected	$\mathbb{D}_{\mathfrak{r}_1} \cup \mathbb{D}_{\mathfrak{r}_2}$
$\mathfrak{r}_1 \to \mathfrak{r}_2$	if $\mathfrak{r}_1$ is being respected, then $\mathfrak{r}_2$ is being respected	$(\mathcal{P}(\mathbb{V})\setminus\mathbb{D}_{\mathfrak{r}_1})\cup(\mathbb{D}_{\mathfrak{r}_1}\cap\mathbb{D}_{\mathfrak{r}_2})$
$\mathfrak{r}_1 \Rightarrow \mathfrak{r}_2^*$	$\mathfrak{r}_1$ is being respected first and selection is carried out, then respect $\mathfrak{r}_2$ given the output $\mathbb{m} \in \mathbb{D}_{\mathfrak{r}_1}$ of the first step	$\{\mathbb{a} : \mathbb{a} \in \mathbb{D}_{\mathfrak{r}_2}, \mathbb{a} \subseteq \mathbb{m}\}$

Selection rule $\mathfrak{r}_i$ on $\mathbb{V}$ is congruent to $\mathbb{D}_{\mathfrak{r}_i}, i = 1, 2$.
* This operation requires the variables potentially being selected by $\mathfrak{r}_1$ are the same with the ones in $\mathfrak{r}_2$.

Table 3.2: Operations for selection rules and the resulting selection dictionaries.

The rule $\mathfrak{r}_{\mathcal{O}}$ resulting from the operation is congruent to a unique selection dictionary which is congruent to $\mathfrak{r}_{\mathcal{O}}$, $\mathbb{D}_{\mathcal{O}}$. We define five operations in Table 3.2. Given an operation on

rules, we can derive the corresponding operation on the related dictionaries that will result in the selection dictionary $\mathbb{D}_{\mathcal{O}}$. Table 3.2 shows the resulting dictionary for each operation. The derivation of each result is given in Appendix A.5. These results allow us to develop algorithms to output selection dictionaries for complex rules through operations on simpler rules.

We use the running example in Table 3.1 to illustrate the second and forth operations on unit rules as a special case.

Define $\mathfrak{u}_1 \coloneqq \mathfrak{u}_{\{1,2\}}(\{A,B\}) \cong \mathbb{D}_{\mathfrak{u}_1} = \big\{\{A\},\{B\},\{A,B\},\{A,C\},\{B,C\},\{A,B,C\}\big\}$,
$\mathfrak{u}_2 \coloneqq \mathfrak{u}_{\{1\}}(\{A\}) \cong \mathbb{D}_{\mathfrak{u}_2} = \big\{\{A\},\{A,B\},\{A,C\},\{A,B,C\}\big\}$,
$\mathfrak{u}_3 \coloneqq \mathfrak{u}_{\{1\}}(\{B\}) \cong \mathbb{D}_{\mathfrak{u}_3} = \big\{\{B\},\{A,B\},\{B,C\},\{A,B,C\}\big\}$.
The $\mathfrak{r}_2$ in Table 3.1 is "if A is selected, then B must be selected," which can be expressed as $\mathfrak{r}_2 \coloneqq \mathfrak{u}_2 \to \mathfrak{u}_3$. According to Table 3.2, $\mathfrak{r}_2$ is congruent to $\big\{\emptyset,\{B\},\{A,B\},\{C\},\{B,C\},\{A,B,C\}\big\}$, which is exactly the $\mathbb{D}_{\mathfrak{r}_2}$ in Table 3.1. Note that, by Definition 1, the operation on two selection rules results in a selection rule, thus the results of an operation on two selection rules can be an input of a second operation. We use parentheses to differentiate the order of operations. The $\mathfrak{r}_3$ in Table 3.1 is "select at least one variable in $\{A,B\}$" $\wedge$ $\mathfrak{r}_2$. Thus, $\mathfrak{r}_3 \coloneqq \mathfrak{u}_1 \wedge (\mathfrak{u}_2 \to \mathfrak{u}_3)$ is a valid operation resulting in a rule that is congruent to the selection dictionary $\big\{\{B\},\{A,B\},\{B,C\},\{A,B,C\}\big\}$ (according to Table 3.2), which is exactly the $\mathbb{D}_{\mathfrak{r}_3}$ in Table 3.1.

Now we use another example to illustrate the last operation. Suppose $\mathbb{V} = \{A,B,C,D\}$, and $\mathfrak{r}_1$ is "$\{A,B\}$ must be selected collectively, same for $\{C,D\}$". That is, $\mathfrak{r}_1 = \mathfrak{u}_{\{0,2\}}(\{A,B\})$ $\wedge\, \mathfrak{u}_{\{0,2\}}(\{C,D\})$. Suppose $\mathfrak{r}_2$ is "if A is selected, then B must be selected, and if C is selected, then D must be selected". That is, $\mathfrak{r}_2 = \big\{\mathfrak{u}_{\{1\}}(\{A\}) \to \mathfrak{u}_{\{1\}}(\{B\})\big\} \wedge \big\{\mathfrak{u}_{\{1\}}(\{C\}) \to \mathfrak{u}_{\{1\}}(\{D\})\big\}$. If the result after respecting $\mathfrak{r}_1$ is $\mathbb{m} = \{A,B\}$, then according to Table 3.2, the dictionary that is congruent to $\mathfrak{r}_1 \Rightarrow \mathfrak{r}_2$ should be $\{\emptyset,\{B\},\{A,B\}\}$. Note that it is not $\mathbb{D}_{\mathfrak{r}_2}$, as $\mathbb{D}_{\mathfrak{r}_2} = \big\{\emptyset,\{B\},\{A,B\},\{D\},\{C,D\},\{B,D\},\{B,C,D\},\{A,B,D\},\{A,B,C,D\},\big\}$

We provide some useful properties of operations below, which can be verified by checking the resulting dictionaries for both sides of the equations. These properties can be used to identify which selection rules are in a same equivalence class.

PROPOSITION 1. *Given* $\mathfrak{r}_1 \neq \mathfrak{r}_2 \neq \mathfrak{r}_3$ *(in the sense that the congruent dictionaries are distinct), then*

1. *Commutative laws:* $\mathfrak{r}_1 \wedge \mathfrak{r}_2 = \mathfrak{r}_2 \wedge \mathfrak{r}_1$*;* $\mathfrak{r}_1 \vee \mathfrak{r}_2 = \mathfrak{r}_2 \vee \mathfrak{r}_1$
2. *Associative laws:* $(\mathfrak{r}_1 \wedge \mathfrak{r}_2) \wedge \mathfrak{r}_3 = \mathfrak{r}_1 \wedge (\mathfrak{r}_2 \wedge \mathfrak{r}_3)$*;* $(\mathfrak{r}_1 \vee \mathfrak{r}_2) \vee \mathfrak{r}_3 = \mathfrak{r}_1 \vee (\mathfrak{r}_2 \vee \mathfrak{r}_3)$
3. *Non-distributive laws:* $\mathfrak{r}_1 \vee (\mathfrak{r}_2 \wedge \mathfrak{r}_3) \neq (\mathfrak{r}_1 \vee \mathfrak{r}_2) \wedge (\mathfrak{r}_1 \vee \mathfrak{r}_3)$*;*
   $\mathfrak{r}_1 \wedge (\mathfrak{r}_2 \vee \mathfrak{r}_3) \neq (\mathfrak{r}_1 \wedge \mathfrak{r}_2) \vee (\mathfrak{r}_1 \wedge \mathfrak{r}_3)$
4. *Sequential laws* $(\mathfrak{r}_1 \to \mathfrak{r}_2) \wedge (\mathfrak{r}_1 \to \mathfrak{r}_3) = \mathfrak{r}_1 \to (\mathfrak{r}_2 \wedge \mathfrak{r}_3)$*;*
   $(\mathfrak{r}_1 \to \mathfrak{r}_2) \vee (\mathfrak{r}_1 \to \mathfrak{r}_3) = \mathfrak{r}_1 \to (\mathfrak{r}_2 \vee \mathfrak{r}_3)$

The next theorem confirms that, equipped with operations and unit rules, we can now effectively express any selection rule in a mathematical language which allows us to develop algorithms to combine multiple rules and generate resulting dictionaries.

THEOREM 3. *All selection rules can be expressed by either unit rules or operations on unit rules using* $\wedge$ *and* $\vee$*.*

The proof is given in Appendix A.6. Next, we use Examples 1 and 2 with a hypothetical data structure to illustrate how to express some common selection dependencies by unit rules and operations. The corresponding selection dictionaries are also provided.

Example 1.1 (Individual selection) In Example 1, suppose all variables are continuous or binary, and no structure is imposed. We can set the selection rule as selection between 0 to 4 variables $\mathfrak{r} = \mathfrak{u}_{\{0,1,2,3,4\}}(\mathbb{V})$, and then $\mathbb{D}_{\mathfrak{r}} = \mathcal{P}(\mathbb{V})$.

Example 1.2 (Groupwise selection) In Example 1, suppose we have 2 three-level categorical variables. Denote $\mathbb{F}_1 = \{A, B\}, \mathbb{F}_2 = \{D, E\}$. Let the variables in $\mathbb{F}_1$ be the dummy variables representing a categorical variable, and similarly for the variables in $\mathbb{F}_2$. In an analysis, we would like to select $\mathbb{F}_1$ collectively, same for $\mathbb{F}_2$. We can then set $\mathfrak{r} = \mathfrak{u}_{\{0,2\}}(\mathbb{F}_1) \wedge \mathfrak{u}_{\{0,2\}}(\mathbb{F}_2)$. (Yuan and Lin, 2006) In addition, $\mathbb{D}_\mathfrak{r} = \{\emptyset, \mathbb{F}_1, \mathbb{F}_2, \mathbb{F}_1 \cup \mathbb{F}_2\}$.

Example 1.3 (Within group selection) If variables in $\mathbb{F}_1$ are one group, and $\mathbb{F}_2$ represents a second group, and the goal is to select at least one variable from both groups, (Campbell and Allen, 2017) then we set $\mathfrak{r} = \mathfrak{u}_{\{1,2\}}(\mathbb{F}_1) \wedge \mathfrak{u}_{\{1,2\}}(\mathbb{F}_2)$, meaning there is at least one variable must be selected in $\mathbb{F}_1$ and $\mathbb{F}_2$ respectively. In addition, $\mathbb{D}_\mathfrak{r} = \big\{\{A, C\}, \{B, C\}, \{A, B, C\}, \{A, D\}$, $\{B, D\}, \{A, B, D\}, \{A, C, D\}, \{B, C, D\}, \{A, B, C, D\}\big\}$

Example 1.4 (Sparse group selection) Consider the following selection rule: first select $\mathbb{F}_1$ and/or $\mathbb{F}_2$, or neither, as groups; then select individual variables from the selected group(s) if at least one group is selected. (Breheny and Huang, 2009; Simon et al., 2013) We set $\mathfrak{r} = \big(\mathfrak{u}_{\{0,2\}}(\mathbb{F}_1) \wedge \mathfrak{u}_{\{0,2\}}(\mathbb{F}_2)\big) \Rightarrow \big(\mathfrak{u}_{\{0,1,2\}}(\mathbb{F}_1) \wedge \mathfrak{u}_{\{0,1,2\}}(\mathbb{F}_2)\big)$, meaning we first execute the same variable selection techniques in Example 1.2, and then select individual variables in the set of variables selected by the previous step. The selection rule dictionary depends on the result from the first step. If the result is none, then $\mathbb{D}_\mathfrak{r} = \{\emptyset\}$, if the result is $\mathbb{F}_1$, then $\mathbb{D}_\mathfrak{r} = \{\emptyset, \{A\}, \{B\}, \{A, B\}\}$, similarly if the result is $\mathbb{F}_2$. If the result is $\mathbb{F}_1 \cup \mathbb{F}_2$, then $\mathbb{D}_\mathfrak{r} = \mathcal{P}(\mathbb{V})$.

Example 2.1 (Categorical interaction selection with strong heredity) In Example 2, because $\{B_1, B_2\}$ are dummy variables representing a same categorical variable, so they have to be collectively selected. Similarly for $\{AB_1, AB_2\}$. In addition, there is a common rule that is being applied in interaction selection, which is called strong heredity (Haris et al., 2016a; Lim and Hastie, 2015): "if the interaction is selected, then all of its main terms must be selected". Define $\mathfrak{u}_1 = \mathfrak{u}_{\{0,2\}}\{B_1, B_2\}, \mathfrak{u}_2 = \mathfrak{u}_{\{0,2\}}\{AB_1, AB_2\}, \mathfrak{u}_3 = \mathfrak{u}_{\{1,2\}}\{AB_1, AB_2\}$, $\mathfrak{u}_4 = \mathfrak{u}_{\{3\}}\{A, B_1, B_2\}$. The selection rule $\mathfrak{r} = (\mathfrak{u}_1 \wedge \mathfrak{u}_2) \wedge (\mathfrak{u}_3 \to \mathfrak{u}_4)$ satisfies the common

selection dependencies imposed for categorical interaction selection and strong heredity. In addition, $\mathbb{D}_{\mathfrak{r}} = \{\emptyset, \{A\}, \{B_1, B_2\}, \{A, B_1, B_2\}, \{A, B_1, B_2, AB_1, AB_2\}\}$.

Example 2.2 (Categorical interaction selection with weak heredity) Another common rule that can be applied on interaction selection is weak heredity (Haris et al., 2016a): "if the interaction is selected, then at least one of its main terms must be selected". To satisfy weak heredity, we further define $\mathfrak{u}_5 = \mathfrak{u}_{\{1,3\}}\{A, B_1, B_2\}$. Then with the predefined unit rules in Example 2.1, the selection rule $\mathfrak{r} = (\mathfrak{u}_1 \wedge \mathfrak{u}_2) \wedge (\mathfrak{u}_3 \rightarrow \mathfrak{u}_5)$ satisfy the common selection dependencies imposed for categorical interaction selection and weak heredity. In addition, the corresponding selection dictionary is the union of the $\mathbb{D}_{\mathfrak{r}}$ in Example 2.1 and $\{\{A, AB_1, AB_2\}, \{B_1, B_2, AB_1, AB_2\}\}$.

3.4 Penalization structure and grouping structure identification

Many existing penalized regression methods (i.e. regularization methods) can respect non-trivial selection rules. (Yuan and Lin, 2006; Campbell and Allen, 2017; Simon et al., 2013; Breheny and Huang, 2009; Haris et al., 2016a; Yuan et al., 2011; Jenatton et al., 2011a; Obozinski et al., 2011a) In fact, the different regularization methods were developed in order to respect different types of rules. In this section, we formalize the framework for penalization structures and describe how it connects to the developed mathematical language. For a regularization method that allows for respecting a given selection rule, this work provides guidance on how to assign grouped variables in penalty terms in order to respect a selection rule.

For a given covariate set $\mathbb{V}$ and outcome O, let $D = (\mathbb{V}, O)$ and denote the coefficients of covariates in the penalized regression by $\boldsymbol{\beta}$. Suppose the data are centered at 0, so that we

omit the intercept, the penalized regression solves

$$\min_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}; D) + \Omega(\boldsymbol{\beta}, \boldsymbol{\theta}), \tag{3.1}$$

where $\ell(\boldsymbol{\beta}; D)$ is a convex loss function, $\boldsymbol{\theta}$ is a vector of hyper parameters, and penalty term $\Omega(\boldsymbol{\beta}, \boldsymbol{\theta})$ is a function of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. Different specifications of Ω result in different regularization methods, with different variable selection results. A grouping structure $\mathbb{G} := \{\mathbb{g}_i, i = 1, ..., I\}$ is a set of non-empty subsets $\mathbb{g}_i \subseteq \mathbb{V}$ s.t. $\cup_{i=1}^{I} \mathbb{g}_i = \mathbb{V}$. Let $\boldsymbol{\beta}_{|\mathbb{g}}$ be a vector of the same length as $\boldsymbol{\beta}$ whose coordinates are equal to those of $\boldsymbol{\beta}$ for indices in the set $\mathbb{g}$ and 0 otherwise. $\|\cdot\|_q$ indicates the L^q norm.

Table 3.3 summarizes five types of penalties with some key rules that they can respect. Each regularization method $\mathcal{M}$ has restrictions on which grouping structures are allowed. We say that the grouping structure is not compatible with a regularization method when the grouping structure does not satisfy the method's restrictions. For example, when some groups contain more than one element, i.e. $|\mathbb{g}| > 1$, for any $\mathbb{g} \in \mathbb{G}$, the grouping structure $\mathbb{G}$ is not compatible with Lasso or Adaptive Lasso.

Method, $\mathcal{M}$	$\Omega(\boldsymbol{\beta}; \boldsymbol{\theta})$	**Condition on $\mathbb{g}$**	**Key rules**
Lasso	$\lambda \sum_{\mathbb{g} \in \mathbb{G}} \left\|\boldsymbol{\beta}_{\|\mathbb{g}}\right\|_1$	$\|\mathbb{g}\| = 1, \forall \mathbb{g}$	$\mathfrak{u}_{\{0,...,\|\mathbb{V}\|\}}(\mathbb{V})$
Adaptive Lasso	$\lambda \sum_{\mathbb{g} \in \mathbb{G}} \omega_{\mathbb{g}} \left\|\boldsymbol{\beta}_{\|\mathbb{g}}\right\|_1$	$\|\mathbb{g}\| = 1, \forall \mathbb{g}$	$\mathfrak{u}_{\{0,...,\|\mathbb{V}\|\}}(\mathbb{V})$
Group Lasso	$\lambda \sum_{\mathbb{g} \in \mathbb{G}} \sqrt{\|\mathbb{g}\|} \left\|\boldsymbol{\beta}_{\|\mathbb{g}}\right\|_2$	$\mathbb{g}_i \cap \mathbb{g}_{i'} = \emptyset, i \neq i', \forall \mathbb{g}$	$\wedge_i \mathfrak{u}_{\{0,\|\mathbb{g}_i\|\}}(\mathbb{g}_i)$
Exclusive Lasso	$\lambda \sum_{\mathbb{g} \in \mathbb{G}} \left\|\boldsymbol{\beta}_{\|\mathbb{g}}\right\|_1^2$	$\mathbb{g}_i \cap \mathbb{g}_{i'} = \emptyset, i \neq i', \forall \mathbb{g}$	$\wedge_i \mathfrak{u}_{\{1,...,\|\mathbb{g}_i\|\}}(\mathbb{g}_i)$
SGL	$(1-\gamma)\lambda \sum_{\mathbb{g} \in \mathbb{G}} \sqrt{\|\mathbb{g}\|} \left\|\boldsymbol{\beta}_{\|\mathbb{g}}\right\|_2 + \gamma\lambda \left\|\boldsymbol{\beta}\right\|_1$	$\mathbb{g}_i \cap \mathbb{g}_{i'} = \emptyset, i \neq i', \forall \mathbb{g}$	$[\wedge_i \mathfrak{u}_{\{0,\|\mathbb{g}_i\|\}}(\mathbb{g}_i)] \Rightarrow [\wedge_i \mathfrak{u}_{\{0,...,\|\mathbb{g}_i\|\}}(\mathbb{g}_i)]$
LOG	$\lambda \sum_{\mathbb{g} \in \mathbb{G}} \omega_{\mathbb{g}} \left\|\boldsymbol{\alpha}_{\|\mathbb{g}}\right\|_2^*$	NA	$\wedge_{i,j} \left[\mathfrak{u}_{\mathbb{C}_i}(\mathbb{g}_i) \rightarrow \mathfrak{u}_{\mathbb{C}_j}(\mathbb{g}_j)\right]$

* $\sum_{\mathbb{g} \in \mathbb{G}} \boldsymbol{\alpha}_{|\mathbb{g}} = \boldsymbol{\beta}$
SGL: Sparse group Lasso, LOG: latent overlapping group Lasso

Table 3.3: Summary of some penalization methods

Given a selection rule and some existing regularization method that could potentially satisfy the selection rule, our framework aims to illustrate how to specify $\boldsymbol{\beta}_{|\mathbb{g}}$, i.e., how to group variables in the penalty term Ω. The framework can be generalized to accommodate more complicated penalties.

We first characterize a "penalization structure" according to a given method and grouping of variables in the penalty term.

DEFINITION 8 (Grouping structure and penalization structure). *For a given $\mathbb{V}$, a **penalization structure** consists of a **grouping structure** and a compatible regularization method $\mathcal{M}$. A grouping structure $\mathbb{G} \coloneqq \{\mathbb{g}_i\}$ s.t. $\cup_{i=1}^{I} \mathbb{g}_i = \mathbb{V}$ is any collection of non-empty subsets of $\mathbb{V}$, whose union is $\mathbb{V}$.*

A grouping structure defines how to assign variables into various groups in the penalty term Ω. The objective of pairing a regularization method with a grouping structure is to implement restrictions on the combinations of variables that can be selected together (which corresponds to respecting a selection rule). The actual variables selected in an analysis will depend on the data, D. Similar to the selection rule dictionary, we define the penalization structure dictionary below.

DEFINITION 9 (Penalization structure dictionary). *Given a method $\mathcal{M}$, and a compatible grouping structure $\mathbb{G}$ on $\mathbb{V}$, there is one corresponding **penalization structure dictionary** $\mathbb{D}_{\mathbb{G}}^{\mathcal{M}}$, which is a dictionary that contains all subsets of $\mathbb{V}$ that could potentially result from the application of the penalization structure on sample data.*

Similar to the relationship between selection rules and selection rule dictionaries, we say the resulting penalization structure dictionary is congruent to the penalization structure, denoted by $\{\mathcal{M}, \mathbb{G}\} \cong \mathbb{D}_{\mathbb{G}}^{\mathcal{M}}$. When the penalization structure dictionary $\mathbb{D}_{\mathbb{G}}^{\mathcal{M}}$ equals a selection

dictionary $\mathbb{D}_{\mathfrak{r}}$ of a rule $\mathfrak{r}$, we say that the penalization structure $\{\mathcal{M}, \mathbb{G}\}$ respects the selection rule $\mathfrak{r}$. That is, when $\{\mathcal{M}, \mathbb{G}\} \cong \mathbb{D}_{\mathbb{G}}^{\mathcal{M}} = \mathbb{D}_{\mathfrak{r}} \cong \mathfrak{r}$, the penalization structure $\{\mathcal{M}, \mathbb{G}\}$ respects (or is congruent to) the selection rule $\mathfrak{r}$.

The grouping structure can be shown graphically, where the variables in a same group $\mathfrak{g}$ are in the same closed curve. We use the previous examples to illustrate various grouping structures with different regularization methods. The related selection rule is given in each example.

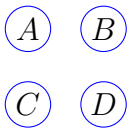


Figure 3.3: Grouping structure that is compatible with Lasso and Adaptive Lasso

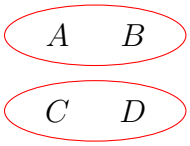


Figure 3.4: Grouping structure that is compatible with Group Lasso

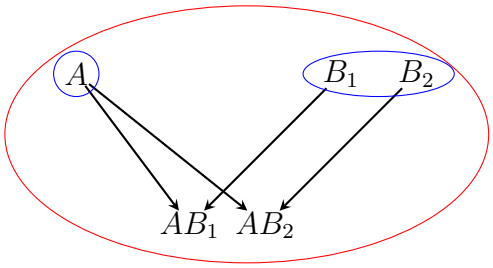


Figure 3.5: Grouping structure that is compatible with latent overlapping group Lasso

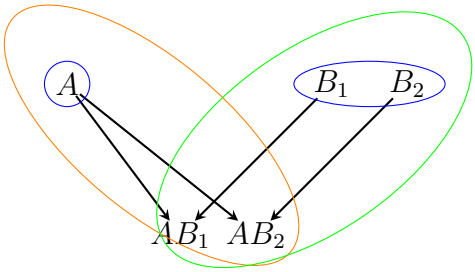


Figure 3.6: Grouping structure that is compatible with latent overlapping group Lasso

Example 1.5 ([Adaptive] Lasso) In Example 1, we can use Lasso (Tibshirani, 1996) or Adaptive Lasso (Zou, 2006) to respect the selection rule in Example 1.1: $\mathfrak{r} = \mathfrak{u}_{\{0,1,2,3,4\}}(\mathbb{V})$ with $\mathbb{V} = \{A, B, C, D\}$. Then the grouping structure given in Figure 3.3 is $\mathbb{G} = \big\{\{A\}, \{B\}, \{C\}, \{D\}\big\}$.

Example 1.6 (Group Lasso) Group Lasso (Yuan and Lin, 2006) can achieve the groupwise selection described in Example 1.2: $\mathfrak{r} = \mathfrak{u}_{\{0,2\}}(\mathbb{F}_1) \wedge \mathfrak{u}_{\{0,2\}}(\mathbb{F}_2)$, where $\mathbb{F}_1 = \{A, B\}$, and $\mathbb{F}_1 = \{C, D\}$. The corresponding grouping structure in Figure 3.4 is $\mathbb{G} = \big\{\mathbb{F}_1, \mathbb{F}_2\big\}$.

Example 1.7 (Exclusive Lasso) Exclusive Lasso (Zhou et al., 2010; Campbell and Allen,

2017) can achieve the within group selection in Example 1.3: $\mathfrak{r} = \mathfrak{u}_{\{1,2\}}(\mathbb{F}_1) \wedge \mathfrak{u}_{\{1,2\}}(\mathbb{F}_2)$. The corresponding grouping structure in Figure 3.4 is $\mathbb{G} = \{\mathbb{F}_1, \mathbb{F}_2\}$.

Example 1.8 (Sparse group Lasso) Sparse group Lasso (Simon et al., 2013) can achieve the sparse group selection in Example 1.5: $\mathfrak{r} = \big(\mathfrak{u}_{\{0,2\}}(\mathbb{F}_1) \wedge \mathfrak{u}_{\{0,2\}}(\mathbb{F}_2)\big) \Rightarrow \big(\mathfrak{u}_{\{0,1,2\}}(\mathbb{F}_1) \wedge \mathfrak{u}_{\{0,1,2\}}(\mathbb{F}_2)\big)$. The corresponding grouping structure in Figure 3.4 is $\mathbb{G} = \{\mathbb{F}_1, \mathbb{F}_2\}$.

Example 2.3 (Latent overlapping group Lasso) Latent overlapping group Lasso (Obozinski et al., 2011a) can achieve categorical interaction selection with strong heredity in Example 2.1 (Lim and Hastie, 2015): $\big(\mathfrak{u}_{\{0,2\}}\{B_1, B_2\} \wedge \mathfrak{u}_{\{0,2\}}\{AB_1, AB_2\}\big) \wedge \big(\mathfrak{u}_{\{1,2\}}\{AB_1, AB_2\} \rightarrow \mathfrak{u}_{\{3\}}\{A, B_1, B_2\}\big)$. The corresponding grouping structure $\mathbb{G} = \big\{\{A\}, \{B_1, B_2\}, \{A, B_1, B_2, AB_1, AB_2\}\big\}$ is shown in Figure 3.5. When a variable appears in more than one group, for example, A in both the first and third group, we say these two groups *overlap*.

Example 2.4 (Latent overlapping group Lasso) The method can also achieve categorical interaction selection with weak heredity in Example 2.2: $\big(\mathfrak{u}_{\{0,2\}}\{B_1, B_2\} \wedge \mathfrak{u}_{\{0,2\}}\{AB_1, AB_2\}\big) \wedge \big(\mathfrak{u}_{\{1,2\}}\{AB_1, AB_2\} \rightarrow \mathfrak{u}_{\{1,3\}}\{A, B_1, B_2\}\big)$. The corresponding grouping structure $\mathbb{G} = \big\{\{A\}, \{B_1, B_2\}, \{A, B_1, B_2, AB_1, AB_2\}, \{A, AB_1, AB_2\}, \{B_1, B_2, AB_1, AB_2, \}\big\}$ is shown in Figure 3.6.

In Example 1.5, we see that different penalization structures can respect the same selection rule. In Examples 1.7 and 1.8, we see that a same grouping structure with different regularization methods can respect different selection rules. In Examples 2.3 and 2.4, we see that with different grouping structures, the same regularization method can achieve various selection rules. However, one penalization structure can only respect one selection rule.

For a given regularization method, it is also possible to establish sufficient and necessary conditions under which grouping structures satisfy selection rules. We focus on latent overlapping group Lasso which can respect many different types of selection rules. The proof of Theorem 4 is given in Appendix A.7.

THEOREM 4. *For a given* $\mathbb{V}$*, with latent overlapping group Lasso, the sufficient and necessary condition of a grouping structure* $\mathbb{G} \coloneqq \{\mathbb{g}_i, i = 1, \ldots, I\}$ *to be congruent to a selection rule* $\mathfrak{r}$ *is* $\mathbb{D}_\mathfrak{r} = \{\cup_{\mathbb{g}\in\mathbb{Q}_j}\mathbb{g}, j = 1, \ldots, 2^I\}$*, where* $\mathbb{Q}_j, j = 1, ..., 2^I$ *are all unique subsets of* $\mathbb{G}$*.*

With latent overlapping group Lasso, the penalization structure dictionary is a set of sets, where each set is a union of groups in a subset of $\mathbb{G}$. When the penalization structure dictionary is the same as the selection dictionary that is congruent to the desired selection rule, the penalization structure is congruent to the selection rule. Theorem 4 provides us with an equation to check if latent overlapping group Lasso can respect a given selection rule, and if so, whether a postulated grouping structure respects the selection rule. Similar theorems can be developed for other regularization methods allowing for grouping structures. For instance the corresponding theorem for overlapping group Lasso is given in Appendix A.8. The proof is similar to the one in Appendix A.7.

Even though Theorem 4 does not tell us how to define a grouping structure for a selection rule, it inspires us how to postulate a grouping structure: starting from the selection rule dictionary, we seek all groups that satisfy the condition. However, it remains a non-trivial task to actually construct this for complex selection rules.

3.5 Selection rule-based variable selection via optimization

While penalized regression methods can be used for variable selection, they cannot control the exact number of variables being selected. This limits their ability to respect certain selection rules. In this section, we use the theoretical framework developed in Section 3.3 to define new optimization techniques that can respect any selection rule that can be represented by a binary operation on unit rules.

Inspired by the optimization technique for best subset selection (Bertsimas et al., 2016; Bertsimas and King, 2017), we propose to use the L^0 norm to directly control the number of variables selected from a specified set. The L^0 norm of a vector counts the non-zero elements in the vector. In a variable selection context, if $\boldsymbol{\beta}$ represents the vector of covariate coefficients, $\|\boldsymbol{\beta}\|_0$ is the number of variables selected. The optimization problem for the best subset selection can then be formulated as

$$\min_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}; D) \qquad \text{s.t.} \quad \|\boldsymbol{\beta}\|_0 \leqslant k. \tag{3.2}$$

where $k \leqslant |\mathbb{V}|$ is a tuning parameter that controls the maximum number of selected variables. Unlike the L^1, L^2, and L^∞ norms that penalize by coefficient magnitude, this norm penalizes the number of variables being selected. Best subset selection corresponds to the same trivial selection rule as Lasso, i.e. $\mathbf{u}_{\{0,\ldots,|\mathbb{V}|\}}(\mathbb{V})$.

Equation (3.2) can be reformulated to a MIO problem (Bertsimas and Weismantel, 2005)

$$\min_{\boldsymbol{\beta},\mathbf{z}} \ell(\boldsymbol{\beta}; D) \quad \text{s.t.} \quad -M\mathbf{z} \leqslant \boldsymbol{\beta} \leqslant M\mathbf{z};\ \mathbf{1}^\intercal\mathbf{z} \leqslant k, \tag{3.3}$$

where $\mathbf{z} \in \{0,1\}^{|\mathbb{V}|}$ and $M = \left\|\hat{\boldsymbol{\beta}}\right\|_\infty$, where $\hat{\boldsymbol{\beta}}$ is the solution of (3.3). The tuning parameter M is chosen to be large enough so that the solution to (3.2) equals the solution to (3.3).(Bertsimas et al., 2016) Decision variable $\mathbf{z}$ corresponds to a binary vector that indicates which variables from $\mathbb{V}$ are selected.

When $\ell(\boldsymbol{\beta}, D)$ is a quadratic loss the above amounts to a mixed-integer quadratic program problem (Lazimy, 1982). The algorithm to solve (3.3) is based on a discrete extension of modern first-order continuous optimization methods, and it can provide near-optimal solutions for the best subset problem in high-dimensional data. (Bertsimas et al., 2016) The algorithm can be implemented by the well-known, powerful heuristic solvers `Gurobi` (Gurobi Optimization, LLC, 2022). The authors also showed via simulation that the method is supe-

rior in identifying the true model covariates relative to Lasso (Tibshirani, 1996), Sparsenet (a family of concave penalties) (Mazumder et al., 2011) and the stand-alone discrete first-order method. More generally, when $\ell(\boldsymbol{\beta}, D)$ is convex but non-differentiable, one can use proximal gradient descent algorithm (Moreau, 1962; Nesterov, 2013b) where each iteration solves an MIO.

While the best subset selection controls the number of selected covariates; the unit rule controls the number of variables being selected from a subset of candidate variables. The best subset selection can be viewed as searching for the best model within the space of all $2^{|\mathbb{V}|}$ possible candidate model; incorporation of the selection rule shrinks the space of candidate models. We first formulate the optimization problem corresponding to an arbitrary unit rule.

1. A unit rule restricts the number of variables being selected in a subset of candidate variables, and does not restrict the others. To respect a unit rule $\mathfrak{u}_{\mathbb{C}}(\mathbb{F})$, we wish to solve

$$\min_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}; D) \qquad \text{s.t.} \quad \left(\left\|\boldsymbol{\beta}_{|\mathbb{F}}\right\|_0, \left\|\boldsymbol{\beta}_{|\mathbb{V}\setminus\mathbb{F}}\right\|_0\right) \in \mathbb{C} \times \mathbb{N}_{\leqslant k},$$

   where $\times$ is the Cartesian product of two sets, and $\mathbb{N}_{\leqslant k} = \{0, \ldots, k\}$ is a set that contains all non-negative integers that are less than or equal to k. This is equivalent to the reformulation of solving $\{\text{Opt}_c : c \in \mathbb{C}\}$, where

$$\text{Opt}_c: \quad \min_{\boldsymbol{\beta}, \mathbf{z}} \ell(\boldsymbol{\beta}; D) \qquad \text{s.t.} \quad -M\mathbf{z} \leqslant \boldsymbol{\beta} \leqslant M\mathbf{z};\ \mathbf{1}^\intercal \mathbf{z}_{|\mathbb{F}} = c;\ \mathbf{1}^\intercal \mathbf{z}_{|\mathbb{V}\setminus\mathbb{F}} \leqslant k.$$

   For a convex loss function ℓ, each Opt_c is a convex MIO with linear constraints and thus can be solved by optimization solvers. We can solve $|\mathbb{C}|$ distinct problems in parallel, and then compare the values of the $|\mathbb{C}|$ objective functions fitted with their local minimizers (the estimated $\boldsymbol{\beta}$ such that the Opt_c is minimal). Then the final

solution of $\{\text{Opt}_c : c \in \mathbb{C}\}$ would be the the minimizer such that $\ell(\boldsymbol{\beta}, D)$ reaches its global minimum, for all $c \in \mathbb{C}$.

We can also define optimization problems corresponding to the binary operations defined in Table 3.2 applied to unit rules.

2. The optimization problem for the selection rule $\neg\mathfrak{u}_{\mathbb{C}}(\mathbb{F})$ parallels the one for the unit rule. We can solve

$$\min_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}; D) \qquad \text{s.t.} \quad \left(\left\|\boldsymbol{\beta}_{|\mathbb{F}}\right\|_0, \left\|\boldsymbol{\beta}_{|\mathbb{V}\setminus\mathbb{F}|}\right\|_0\right) \in (\mathbb{N}_{\leqslant|\mathbb{F}} \setminus \mathbb{C}) \times \mathbb{N}_{\leqslant k}.$$

   This is equivalent to $\{\text{Opt}_c : c \in \mathbb{N}_{\leqslant|\mathbb{F}|} \setminus \mathbb{C}\}$. In this case, we solve $|\mathbb{N}_{\leqslant|\mathbb{F}|} \setminus \mathbb{C}|$ distinct MIOs in parallel.

3. For the selection rule $\mathfrak{u}_{\mathbb{C}_1}(\mathbb{F}_1) \wedge \mathfrak{u}_{\mathbb{C}_2}(\mathbb{F}_2)$, we need to control the numbers of variables being selected in two subsets, and the rest of the variables are unrestricted. We then solve

$$\min_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}; D) \qquad \text{s.t.} \quad \left(\left\|\boldsymbol{\beta}_{|\mathbb{F}_1}\right\|_0, \left\|\boldsymbol{\beta}_{|\mathbb{F}_2}\right\|_0, \left\|\boldsymbol{\beta}_{|\mathbb{V}\setminus(\mathbb{F}_1\cup\mathbb{F}_2)}\right\|_0\right) \in \mathbb{C}_1 \times \mathbb{C}_2 \times \mathbb{N}_{\leqslant k},$$

   which is equivalent to $\{\text{Opt}_{c_1,c_2} : (c_1, c_2) \in \mathbb{C}_1 \times \mathbb{C}_2\}$, where

$$\text{Opt}_{c_1,c_2} : \min_{\boldsymbol{\beta},\mathbf{z}} \ell(\boldsymbol{\beta}; D) \qquad \text{s.t.} \quad -M\mathbf{z} \leqslant \boldsymbol{\beta} \leqslant M\mathbf{z}; \mathbf{1}^\intercal\mathbf{z}_{|\mathbb{F}_1} = c_1;$$
$$\mathbf{1}^\intercal\mathbf{z}_{|\mathbb{F}_2} = c_2; \mathbf{1}^\intercal\mathbf{z}_{|\mathbb{V}\setminus(\mathbb{F}_1\cup\mathbb{F}_2)} \leqslant k,$$

   which requires solving $|\mathbb{C}_1| \times |\mathbb{C}_2|$ distinct MIOs in parallel.

4. For the selection rule $\mathfrak{u}_{\mathbb{C}_1}(\mathbb{F}_1) \vee \mathfrak{u}_{\mathbb{C}_2}(\mathbb{F}_2)$, we need to satisfy at least one unit rule, which

solves

$$\min_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}; D) \qquad \text{s.t.} \quad \left(\left\|\boldsymbol{\beta}_{|\mathbb{F}_1}\right\|_0, \left\|\boldsymbol{\beta}_{|\mathbb{V}\setminus\mathbb{F}_1}\right\|_0\right) \in \mathbb{C}_1 \times \mathbb{N}_{\leqslant k_1} \quad \text{or}$$
$$\left(\left\|\boldsymbol{\beta}_{|\mathbb{F}_2}\right\|_0, \left\|\boldsymbol{\beta}_{|\mathbb{V}\setminus\mathbb{F}_2}\right\|_0\right) \in \mathbb{C}_2 \times \mathbb{N}_{\leqslant k_2},$$

which is equivalent to the optimization problem of finding the minimizer of $\{\text{Opt}^1_{c_1} : c_1 \in \mathbb{C}_1\}$ and $\{\text{Opt}^2_{c_2} : c_2 \in \mathbb{C}_2\}$, where

$$\text{Opt}^i_{c_i}: \quad \min_{\boldsymbol{\beta},\mathbf{z}} \ell(\boldsymbol{\beta}; D) \qquad \text{s.t.} \quad -M\mathbf{z} \leqslant \boldsymbol{\beta} \leqslant M\mathbf{z};\ \mathbf{1}^\intercal \mathbf{z}_{|\mathbb{F}_i} = c_i;\ \mathbf{1}^\intercal \mathbf{z}_{|\mathbb{V}\setminus\mathbb{F}_i} \leqslant k_i,$$

where $i = 1, 2$. It requires solving $|\mathbb{C}_1| + |\mathbb{C}_2|$ distinct MIOs in parallel with two tuning parameters, k_1 and k_2.

5. For the selection rule $\mathfrak{u}_{\mathbb{C}_1}(\mathbb{F}_1) \to \mathfrak{u}_{\mathbb{C}_2}(\mathbb{F}_2)$, it is equivalent to say if $\mathfrak{u}_{\mathbb{C}_1}(\mathbb{F}_1)$ is respected, then respect $\mathfrak{u}_{\mathbb{C}_2}(\mathbb{F}_2)$, if $\mathfrak{u}_{\mathbb{C}_1}(\mathbb{F}_1)$ is not respected, then we do not impose a constraint on $\mathbb{F}_2$. That is, $\mathfrak{u}_{\mathbb{C}_1}(\mathbb{F}_1) \to \mathfrak{u}_{\mathbb{C}_2}(\mathbb{F}_2) = \{\mathfrak{u}_{\mathbb{C}_1}(\mathbb{F}_1) \wedge \mathfrak{u}_{\mathbb{C}_2}(\mathbb{F}_2)\} \vee \{\neg\mathfrak{u}_{\mathbb{C}_1}(\mathbb{F}_1)\}$. Thus we wish to solve

$$\min_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}; D) \qquad \text{s.t.} \qquad \left(\left\|\boldsymbol{\beta}_{|\mathbb{F}_1}\right\|_0, \left\|\boldsymbol{\beta}_{|\mathbb{F}_2}\right\|_0, \left\|\boldsymbol{\beta}_{|\mathbb{V}\setminus(\mathbb{F}_1\cup\mathbb{F}_2)}\right\|_0\right) \in \mathbb{C}_1 \times \mathbb{C}_2 \times \mathbb{N}_{\leqslant k_1} \qquad \text{or}$$
$$\left(\left\|\boldsymbol{\beta}_{|\mathbb{F}_1}\right\|_0, \left\|\boldsymbol{\beta}_{|\mathbb{V}\setminus\mathbb{F}_1}\right\|_0\right) \in (\mathbb{N}_{\leqslant|\mathbb{F}_1|} \setminus \mathbb{C}_1) \times \mathbb{N}_{\leqslant k_2},$$

It can be reformulated the optimization problem of finding the minimizer of $\{\text{Opt}^1_{c_1,c_2} : (c_1, c_2) \in \mathbb{C}_1 \times \mathbb{C}_2\}$ and $\{\text{Opt}^2_{c_3} : c_3 \in \mathbb{N}_{\leqslant|\mathbb{F}_1|} \setminus \mathbb{C}_1\}$, where

$$\text{Opt}^1_{c_1,c_2}: \min_{\boldsymbol{\beta},\mathbf{z}} \ell(\boldsymbol{\beta}; D) \qquad \text{s.t.} \quad -M\mathbf{z} \leqslant \boldsymbol{\beta} \leqslant M\mathbf{z}; \mathbf{1}^\intercal \mathbf{z}_{|\mathbb{F}_1} = c_1;$$
$$\mathbf{1}^\intercal \mathbf{z}_{|\mathbb{F}_2} = c_2; \mathbf{1}^\intercal \mathbf{z}_{|\mathbb{V}\setminus(\mathbb{F}_1\cup\mathbb{F}_2)} \leqslant k_1,$$
$$\text{Opt}^2_{c_3}: \min_{\boldsymbol{\beta},\mathbf{z}} \ell(\boldsymbol{\beta}; D) \qquad \text{s.t.} \quad -M\mathbf{z} \leqslant \boldsymbol{\beta} \leqslant M\mathbf{z};\ \mathbf{1}^\intercal \mathbf{z}_{|\mathbb{F}_1} = c_3;\ \mathbf{1}^\intercal \mathbf{z}_{|\mathbb{V}\setminus\mathbb{F}_1} \leqslant k_2,$$

Both of $\text{Opt}^1_{c_1,c_2}$ and $\text{Opt}^2_{c_3}$ are MIOs. Similar to the above procedures, by solving $|\mathbb{C}_1| \times |\mathbb{C}_2|$ times of $\text{Opt}^1_{c_1,c_2}$, and $|\mathbb{N}_{\leqslant|\mathbb{F}_1|} \setminus \mathbb{C}_1|$ times of $\text{Opt}^2_{c_3}$ in parallel and comparing all the objective functions fitted with their local minimizers, one can thus find the global minimizer.

While this provides an interesting and potentially useful application of our framework, the above development is limited to binary operations on unit rules, which does not cover all possible selection rules. Though we provide the optimization formulations for unit rules and their operations, the theoretical properties of the estimators are not investigated. Both of these topics merit further research.

3.6 Discussion

Structured variable selection can improve model interpretability and also prediction accuracy. Past work has focused on respecting a specific class of selection rules. Our contribution in this paper is the development of a mathematical framework for structured variable selection in full generality, allowing for the incorporation any a priori knowledge about covariate structure into the variable selection to arrive at an interpretable selected model.

Our framework allows for a universal formulation of a priori selection structures using a mathematical language. We then presented a bridge between an arbitrary rule and the related selection dictionary, which is the space of all allowable covariate subsets. The formula that allows for the derivation of the selection dictionary is useful even in a low-dimensional covariate setting since manually listing the selection dictionary requires a great amount of work and is error-prone. The properties and relationships of the defined mathematical objects were also investigated, which was helpful to understand the framework and identify potential future development.

In addition, we established that the selection dictionary is the key to connecting the selection

rule to the implementation of a penalized regression method, which can identify the grouping structure. More importantly, we show in Section 3.5 that the framework can be used to develop new optimization techniques to satisfy a wider spectrum of selection rules.

Our framework unifies the structured variable selection problem, and creates a paradigm where researchers can view the problem generically, rather than starting from a specific class of covariate structure. Generic guidance for variable selection rules would allow practitioners to scrutinize the covariate structures in their application carefully and potentially incorporate a larger scope of desirable selection rules. Given the increasing complexity of data and applications, if new types of selection rules emerge in the future, we expect that our framework will be able to incorporate them.

The framework is helpful in recognizing the scope of selection rules that a method can respect, and understanding the reasons. Theorems 4 and 5 provide conditions for a grouping structure to respect a selection rule in (latent) overlapping group Lasso, which are examples of applying the framework to penalized regression. We know that though the two methods can respect many types of selection rules, it is not always possible – in such a case, no grouping structure can satisfy the equation in theorem 4. For instance, given $\mathbb{V} = \{A, B, C\}$, latent overlapping group Lasso cannot respect the unit rule $\mathfrak{u}_{\{2\}}(\{A, B, C\})$ (with resulting unit dictionary $\big\{\{A, B\}, \{B, C\}, \{A, C\}\big\}$). The reason follows: from Theorem 4, $\{\emptyset\}$ is a subset of $\mathbb{G}$, as well as the $\mathbb{G}$ itself. Therefore, the penalization structure dictionary always contain $\{\emptyset\}$ and the universal set (of $\mathbb{V}$, since the union of all elements in $\mathbb{G}$ must be $\mathbb{V}$ according to the definition of grouping structure). That is, any selection dictionary that not includes these two sets would never equals the penalization structure dictionary when the method used is the latent overlapping group Lasso (similar for the overlapping group Lasso, see Theorem 5). In other words, (latent) overlapping group Lasso can only respect a unit rule when the $\mathbb{C}$ contains both 0 and $|\mathbb{F}|$. In addition, from Table 3.2, the selection dictionary obtained from some operations $(\wedge, \vee, \rightarrow)$ of such unit dictionaries still contain

these two sets. Therefore, the (latent) overlapping grouping Lasso also cannot respect the selection rules that operated from such unit rules.

Ongoing work involves leveraging the theorems to provide roadmaps for the construction of the appropriate grouping structure to respect different types of selection rules. Further conditions and roadmaps for other penalized regression methods are interesting future directions.

We show in section 3.5 that variable selection under unit rules and binary operations on these can be formulated as MIOs, which can be solved by optimization solvers. Though the associated statistical properties are not investigated here, this opens up a research area in structured variable selection , including selection rule classes that cannot be respected by penalized regression methods. It reverses the perspective of the implementation of variable selection methods – rather than starting from existing methods that can respect certain types of selection rules, users can start from their a priori knowledge, set the desired selection rules, and then identify or set up the corresponding methods.

Acknowledgment

The authors would like to thank Alexander William Levis for the valuable discussion. GW is supported by Fonds de Recherche du Québec—Santé Doctoral Training Award (272161), the Faculté de Pharmacie, Université de Montréal through MES, and a CIHR Foundation Grant (FDN-143297) through RWP. ME. Schnitzer is supported by a Canadian Institutes of Health Research Canada Research Chair tier 2 and a Natural Sciences and Engineering Research Council of Canada Discovery Grant. RW. Platt is supported by CIHR Foundation Grant (FDN-143297).

Chapter 4

Structured variable selection: an application in identifying predictors of major bleeding among hospitalized hypertensive patients using oral anticoagulants for atrial fibrillation

Preamble to Manuscript 2. In Chapter 3, I developed a new framework for structured variable selection. In this chapter, I use a complex example in pharmacoepidemiology to demonstrate an application of the framework to penalized regression (the latent overlapping group Lasso), give a step-by-step demonstration of how to derive the selection dictionary and group variables, and also give roadmaps for grouping structure identification for some common selection rules. We use a linked dataset extracted from claims and medical services databases to identify predictors of major bleeding among hospitalized hypertensive patients using oral anticoagulants for atrial fibrillation. In this application, we consider the variable

"the proxy of the adherence to anticoagulant medication" and its interactions with dose and type of oral anticoagulants, respectively. We also consider drug-drug interactions. Seven important selection rules are integrated into the variable selection. Unlike the (adaptive) Lasso, which does not incorporate the selection rules, the selected model from the latent overlapping group Lasso with our defined grouping structure respects the combination of all selection rules, and also resulted in a lower cross-validated risk. This manuscript was submitted to *Statistics in Medicine*.

Note that the supplementary material for this chapter can be found in Appendix B.

Structured variable selection: an application in identifying predictors of major bleeding among hospitalized hypertensive patients using oral anticoagulants for atrial fibrillation

Guanbo Wang[1], Mireille E. Schnitzer[2,3,1], Robert W. Platt[4,1], Rui Wang[5,6], Marc Dorais[7], and Sylvie Perreault[2]

[1]*Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, QC, Canada*

[2]*Faculté de pharmacie, Université de Montréal, Montréal, QC, Canada*

[3]*Département de médecine sociale et préventive, ESPUM, Université de Montréal, Montréal, QC, Canada*

[4]*Department of Pediatrics, McGill University, Montréal, QC, Canada*

[5]*Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, MA, USA*

[6]*Harvard T. H. Chan School of Public Health, Harvard University, Boston, MA, USA*

[7]*StatSciences Inc., Notre-Dame-de-l'Île-Perrot, Quebec, Canada*

This chapter contains the corresponding paper to be submitted to *Statistics in Medicine*

Abstract

Predictor identification is important in medical research as it can help clinicians to have a better understanding of disease epidemiology and identify patients at higher risk of an outcome. Variable selection is often used to reduce the dimensionality of a prediction model. When conducting variable selection, it is often beneficial to take selection dependencies into account. Selection dependencies can help to improve model interpretability, increase the chance of recovering the true model, and augment the prediction accuracy of the resulting model. The latent overlapping group lasso can achieve the goal of incorporating some types of selection dependencies into variable selection by assigning coefficients to different groups of penalties. However, when the selection dependencies are complex, there is no roadmap for how to specify the groups of penalties. Wang et al. (2021) proposed a general framework for structured variable selection, and provided a condition to verify whether a penalty grouping respects a set of selection dependencies. Based on this previous work, we construct roadmaps to derive the grouping specification for some common selection dependencies and apply them to the problem of constructing a prediction model for major bleeding among hypertensive patients recently hospitalized for atrial fibrillation and then prescribed oral anticoagulants. In the application, we consider a proxy of adherence to anticoagulant medication and its interaction with dose and oral anticoagulants type, respectively. We also consider drug-drug interactions. Our method allows for algorithmic identification of the grouping specification even under the resulting complex selection dependencies.

4.1 Introduction

In medical research, identifying important predictors for an outcome can help clinicians develop a better understanding of disease epidemiology, i.e., which factors are associated with a higher risk of a disease, and the independent contribution of each one to the overall prediction. (Schooling and Jones, 2018) In addition, the ensuing predictive models can assist clinicians in identifying patients who are at higher risk, in order to potentially provide different care or more intensive follow-up. Predictor identification can also be regarded as model exploration for causal hypotheses: the predictors being identified can be potentially investigated in future causal analyses and randomized controlled trials. (Shmueli, 2010; Kalisch and Bühlman, 2007; Shortreed and Ertefaie, 2017) In this context, constructing a non-parametric predictive model may be less preferable because black box models are difficult to interpret and prediction accuracy is not the only goal that we pursue. To strike a balance between prediction accuracy and interpretability, a vast literature of variable selection techniques under generalized linear models has emerged. (Tibshirani, 1996; Zou, 2006; Yuan and Lin, 2006; Bhatnagar et al., 2020; Breheny and Huang, 2009; Mairal et al., 2010; Jacob et al., 2009)

Selection dependencies are often inherent or desirable in variable selection. For example, all non-reference binary indicators representing a categorical variable should be selected collectively. As a second example, the selection of a variable (for instance, the interaction) can depend on whether other variables (for instance, main terms) are also selected. Such selection dependencies constrain the candidate models: there are a limited number of models, consisting of different combinations of covariates, that can satisfy the selection dependencies. Models that satisfy such selection dependencies ensure the interpretability of the results.

Recently, Wang et al. (2021) developed a framework for structured variable selection that allows practitioners to respect any combination of selection dependencies, called a "*selection*

rule". The framework enables one to express the selection dependencies using a uniform mathematical language, and then construct the corresponding *selection dictionary*, which is a set that contains all subsets of candidate variables that respect the selection rule. We can perform model selection by fitting each candidate model in the selection dictionary, and select the model that has the best performance in some sense (for example, the lowest cross-validated risk under some loss function).

When the covariates considered in an application are of greater dimension, applying this exhaustive method is computationally inefficient. Another solution to accommodate complicated selection dependencies in variable selection is penalized regression, where the selection dependency is achieved by specifying a *grouping structure*, that is, assigning (possibly overlapping) groups of coefficients of covariates in the penalty term. However, how to identify the grouping structure is not trivial. Wang et al. (2021) gave a sufficient and necessary condition for a grouping structure to respect a selection dependency with latent overlapping group lasso (Obozinski et al., 2011b), given a known selection dictionary. However, it does not tell us exactly how identify the grouping structure.

With the increasing use of administrative claims and electronic health records (EHR), higher dimensional patient information is accessible which may allow for more powerful predictive modeling. However, the selection dependencies may be complex in such data. In this work, we apply and demonstrate the practical use of the framework developed in Wang et al. (2021) to identify predictors of major bleeding among hypertensive patients hospitalized for atrial fibrillation and prescribed oral anticoagulants (OACs) after hospital discharge. In this application, a proxy of adherence to OACs and drug-drug interactions are considered. In our modeling, we retain the structure of the data, and take desirable selection dependencies into account. We also provide related roadmaps for grouping structure identification for some common selection rules.

The remainder of the paper is organized as follows. Section 4.2 introduces the data example.

Section 4.3 gives a brief introduction of the framework developed by (Wang et al., 2021), and illustrates how can we apply and extend the framework. The results of the data analyses are given in section 4.4, followed by a discussion in section 4.5.

4.2 Application: identification of predictors for major bleeding in patients taking OACs

Atrial fibrillation is a disease characterized by an irregular heartbeat, which is due to electrical signal disturbances of the heart. Patients with this condition have a higher risk of stroke, heart failure, and other cardiovascular complications. (Lip and Tse, 2007) To decrease the occurrence of stroke, most patients have to take anticoagulants long-term. (Yamashiro et al., 2019) Warfarin (vitamin K antagonist) was the mainstay anticoagulant for non-valvular atrial fibrillation. (Friberg et al., 2012) However, close monitoring of the International Normalized Ratio (INR) and frequent dose adjustments due to numerous drug interactions are required when taking warfarin. (Connolly et al., 2008) In recent decades, DOACs, including Rivaroxaban, Dabigatran, and Apixaban, became available to non-valvular atrial fibrillation patients. As alternatives to warfarin, they do not require such routine monitoring and are given as fixed doses based on patient characteristics. Nevertheless, non-compliance (Obamiro et al., 2016; Garkina et al., 2016), different dose levels, drug interactions (Raccah et al., 2018; Burn and Pirmohamed, 2018) and contraindications (Schnitzer et al., 2020) complicate the usage of anticoagulants. Patients taking different anticoagulants with various doses and adherence patterns may have different risks of major bleeding. Beyond that, many known or unknown factors may also contribute to the variability of outcomes. Therefore, before measuring the effects of anticoagulants, or predicting the risk of major bleeding, we are interested in investigating the predictors that are associated with or predictive of major bleeding through modeling. (Tripodi et al., 2018; Claxton et al., 2018)

Data source and population-based cohort definition

We use the dataset from Qazi et al. (2021), which was compiled from a subset of the Régie de l'Assurance Maladie du Québec (RAMQ) drug and medical services database linked with the Med-Echo hospitalization database using encrypted patient healthcare insurance numbers. (Tamblyn et al., 1995; Perreault et al., 2020; Eguale et al., 2010; Wilchesky et al., 2004) They identified patients hospitalized for any cause and discharged alive in the community from 2011 to 2017 with a primary or secondary diagnosis of atrial fibrillation. Cohort entry (index date) was defined as the time of the first OAC claim.

We are particularly interested in assessing if adherence (high/low) to the prescribed OAC is associated with the risk of major bleeding (Perreault et al., 2019). However, the history of OAC usage is not available because the cohort consists of the patients who used OACs for the first time. We thus used the history of hypertension drug usage, which are also taken chronically, as a proxy of the adherence to OACs. (Sabaté and Sabaté, 2003) We thus limited the cohort to patients with a previous diagnosis of hypertension who were prescribed at least one hypertension drug in the three months prior to the index date. The complete inclusion and exclusion criteria with patient totals are shown in supplementary material 1 Section B.1.

Study outcome, baseline characteristics and predictor candidates

The outcome is defined as incident major bleeding within 1 year of follow-up. Validated ICD-9 and ICD-10 codes for the outcome are given in supplementary material 2. (Villines et al., 2015; Yao et al., 2016; Lauffenburger et al., 2015; Maura et al., 2015; Graham et al., 2015; Go et al., 2017)

We screened the dataset and selected variables that were either known to be predictive of major bleeding or of particular clinical interest, (Qazi et al., 2021; Landefeld and Goldman, 1989; Roldán et al., 2013; Chao et al., 2018) and the availability of data.

When available, we included the variables in the HAS-BLED (Hypertension, Abnormal Renal/Liver Function, Stroke, Bleeding History or Predisposition, Labile INR, Elderly,

Name	Definition
`DOAC`	1 if the patient was prescribed DOAC, 0 if the patient was prescribed warfarin at cohort entry
`Apixaban`	1 if the patient was prescribed Apixaban at cohort entry, 0 otherwise
`Dabigatran`	1 if the patient was prescribed Dabigatran at cohort entry, 0 otherwise
`High-dose-DOAC`	1 if the patient was prescribed high-dose-DOACs at cohort entry, 0 otherwise
`High-adherence`*	1 if the patient's adherence level was greater than or equal to 0.8

*For the complete definition of adherence, see supplementary material 2

Table 4.1: Definitions of variables related to OAC usage

Drugs/Alcohol Concomitantly) chart (Pisters et al., 2010) and known risk factors of bleeding when taking OACs (Lane and Lip, 2012); the definitions of these variables are given in supplementary material 2. The potential predictors in our analysis included age, sex, CHA_2DS_2-VASc score, comorbidities within 3 years before cohort entry, OAC usage at cohort entry, concomitant medication usage within 2 weeks before cohort entry, and the drug-drug interactions between OAC type and concomitant medications. Definitions of the main terms (binary) variables related to OAC usage are given in Table 4.1. `DOAC` indicates that a patient was prescribed a DOAC, with warfarin as the alternative. We also included indicators of Apixaban and Dabigatran prescriptions with Rivaroxaban being the reference DOAC. DOACs can also be prescribed at high or low-doses. Our variable `High-dose-DOAC` indicates that a patient was receiving a high-dose of a DOAC, where `High-dose-DOAC`$= 0$ means that the patient was either receiving low-dose-DOAC or warfarin. Note that the dosing of warfarin is individualized and not considered in this analysis. Adherence was calculated as the number of days of dispensed hypertension drugs divided by the duration of the prescription period prior to the index date. The variable `High-adherence` was defined as $\geqslant 80\%$ adherence to hypertension drugs.

Selection rules

Given these covariates and the interactions of interest, we consider three types of selection rules. We assume strong heredity (Haris et al., 2016b) for all interactions: if the interaction is selected, its main terms must be selected. The rationale for each rule follows.

1. Selection rules for OAC usage:

   1.1. If `High-dose-DOAC` is selected, then `DOAC` must be selected.

   1.2. If `Apixaban` is selected, then `DOAC` must also be selected.

   1.3. If `Dabigatran` is selected, then `DOAC` must also be selected.

   1.4. If the interaction of `DOAC` and `High-adherence` is selected, then both `DOAC` and `High-adherence` must be selected.

   1.5. When the interaction of `High-dose-DOAC` and `High-adherence` is selected, then the model must also include: `DOAC`, `High-adherence`, `High-dose-DOAC` and the interaction of `DOAC` and `High-adherence`.

2. Selection rule for drug-drug interaction: If a drug-drug interaction is selected, both `DOAC` and the other medication must be selected. (Note that we only include drug-drug interactions between `DOAC` and concomitant medications.)

3. Selection rule for pre-selected variables: The following established predictors for major bleeding are forced into the model: 1) `Age`, 2) `Sex`, 3) `Stroke`, 4) `Anemia`, 5) `Malignancy`, 6) `Liver diseases`, 7) `History of major bleeding`, 8) `Renal diseases`, 9) `Antiplatelets`, 10) `NSAIDs`.

Rule 1.1 is needed since when `DOAC` is in the model, and if `High-dose-DOAC` is selected, then the interpretation of the coefficient of `High-dose-DOAC` is the contrast (e.g. log odds ratio) of high-dose-DOAC versus low-dose-DOAC, which is of interest. If `Apixaban` and `Dabigatran` are also in the model, the coefficient of `High-dose-DOAC` represents the contrast between high-dose-Rivaroxaban versus low-dose-Rivaroxaban. However, without `DOAC`

in the model, these relevant interpretations would be lost. Rule 1.2 is needed since when `DOAC` is in the model, and if `Apixaban` is selected, then the interpretation of the coefficient of `Apixaban` is the contrast of Apixaban and Rivaroxaban. If `High-dose-DOAC` is also in the model, the interpretation would be the contrast of low-dose-Apixaban versus low-dose-Rivaroxaban. However, without `DOAC`, the coefficient of `Apixaban` would represent a contrast against warfarin and Rivaroxaban combined, which is less interpretable. The same rationale applies for `Dabigatran` in rule 1.3. The rationale for rule 1.5 is that: 1) according to strong heredity, both `High-dose-DOAC` and `High-adherence` must be in the model when the interaction of `High-dose-DOAC` and `High-adherence` is selected, 2) when `High-dose-DOAC` is in the model, `DOAC` must be selected, which is justified by the rule 1.1, and 3) the interaction of `High-dose-DOAC` and `High-adherence` represents a three-way interaction between `High-dose-DOAC`, `DOAC`, and `High-adherence`. Therefore, by the rationale of strong heredity, we must include the lower order interaction between `DOAC` and `High-adherence`. If we do not, we are assuming that `High-adherence` has the same impact whether a patient takes warfarin or low-dose-DOAC but the impact is different for a high-dose-DOAC. The coefficient of the main term `High-adherence` would then be less interpretable. The rules 1.4 and 2 are straight-forward applications of strong heredity. Rule 3 is based on the findings from Lane and Lip (2012).

4.3 Statistical methods

4.3.1 Selection rule and selection dictionary

We next present the method to derive the selection dictionary for the defined selection rules, which is rooted in the framework developed by Wang et al. (2021). Recall that the "selection dictionary" is the set that contains all subsets of candidate variables that respect a given selection rule.

Denote the set of candidate covariates by $\mathbb{V}$, with non-empty subsets $\mathbb{F}_1$ and $\mathbb{F}_2$. We use

$\mathcal{P}(\mathbb{V})$ to denote the power set of $\mathbb{V}$, which is the set of all subsets of $\mathbb{V}$. This represents the dictionary without any selection rules applied.

1. Define a *unit rule* $\mathfrak{u}(\mathbb{F}_1)$ as "select $\mathbb{F}_1$", meaning that this selection rule forces the set $\mathbb{F}_1$ into the model. The related selection dictionary $\mathbb{D}_{\mathfrak{u}_1}$ (called "*unit dictionary*") contains all sets which are the union of $\mathbb{F}_1$ and any subset of $\mathbb{V} \setminus \mathbb{F}_1$. That is, $\mathbb{D}_{\mathfrak{u}_1} = \{\mathbb{F}_1 \cup \mathbb{m} : \mathbb{m} \in \mathcal{P}(\mathbb{V}) \setminus \mathbb{F}_1\}$. We define $\mathfrak{u}(\mathbb{F}_2)$ and $\mathbb{D}_{\mathfrak{u}_2}$ similarly with respect to $\mathbb{F}_2$.

2. Define an *if-then rule* $\mathfrak{r}$ as "if $\mathbb{F}_1$ is selected, then $\mathbb{F}_2$ is selected", or equivalently, "if the unit rule $\mathfrak{u}(\mathbb{F}_1)$ is respected, then the unit rule $\mathfrak{u}(\mathbb{F}_2)$ must be respected". This is denoted by $\mathfrak{r} = \mathfrak{u}(\mathbb{F}_1) \to \mathfrak{u}(\mathbb{F}_2)$. The selection dictionary corresponding to the if-then rule is given by $\mathbb{D}_{\mathfrak{r}} = \{\mathcal{P}(\mathbb{V}) \setminus \mathbb{D}_{\mathfrak{u}_1}\} \cup (\mathbb{D}_{\mathfrak{u}_1} \cap \mathbb{D}_{\mathfrak{u}_2})$.

We observed that all the selection rules defined in the previous section can be represented by unit rules and if-then rules. We next show how to derive the corresponding selection dictionary that respects all the selection rules using set operations on the rule-specific dictionaries in Algorithm 1. We also show the exhaustive search method in Algorithm 2. Algorithm 1 constructs the selection dictionary by the unit dictionaries and their operations; whereas Algorithm 2 eliminates the sets in $\mathcal{P}(\mathbb{V})$ that do not respect the rules. Since rule 3 forces the selection of variables, they have to be in each set of the selection dictionary. Both algorithms are available in the Github https://github.com/Guanbo-W/SelectionDictionary.

After running the algorithms, we found that the cardinality of the selection dictionary corresponding to the combination of our 7 rules is 32512. Algorithm 2 took 5 minutes while the exhaustive search took 3.5 seconds on a local computer without the use of parallel computing. Note that though the former approach took longer, it is a generic procedure that can be applied to more complex rule operations outside of our application.

Algorithm 1 Steps of deriving the selection dictionary by set operations

Denote the variables in rule 3 by $\mathbb{A}$.

Define $\mathbb{V}$ as a set that contains all candidate variables except $\mathbb{A}$.

Conduct steps 1 and 2 for each if-then rule (1.1-2) $\mathfrak{r}_j = \mathfrak{u}_{j,1}(\mathbb{F}_{j,1}) \rightarrow \mathfrak{u}_{j,2}(\mathbb{F}_{j,2})$:

Step 1: Derive the unit dictionaries $\mathbb{D}_{\mathfrak{u}_{j,1}} = \{\mathbb{F}_{j,1} \cup \mathbb{m} : \mathbb{m} \in \mathcal{P}(\mathbb{V}) \setminus \mathbb{F}_1\}$. Similarly for $\mathbb{D}_{\mathfrak{u}_{j,2}}$.

Step 2: Derive the selection dictionary of $\mathfrak{r}_j$ as $\mathbb{D}_{\mathfrak{r}_j} = \{\mathcal{P}(\mathbb{V}) \setminus \mathbb{D}_{\mathfrak{u}_{j,1}}\} \cup (\mathbb{D}_{\mathfrak{u}_{j,1}} \cap \mathbb{D}_{\mathfrak{u}_{j,2}})$.

Step 3: Derive the selection dictionary that respects all rules 1.1-2 by $\mathbb{D}_{\mathfrak{r}} = \cap_j \mathbb{D}_{\mathfrak{r}_j}$.

Step 4: The final selection dictionary is $\{\mathbb{A} \cup \mathbb{n} : \mathbb{n} \in \mathbb{D}_{\mathfrak{r}}\}$.

Algorithm 2 Steps of deriving the selection dictionary by exhaustive search

Denote the variables in rule 3 by $\mathbb{A}$.

Define $\mathbb{V}$ as a set containing all candidate variables except $\mathbb{A}$.

For each subset of $\mathcal{P}(\mathbb{V})$, denoted $\mathbb{x}$:

Step 1: for all if-then rules (1.1-2) denoted $\mathfrak{r}_j = \mathfrak{u}_{j,1}(\mathbb{F}_{j,1}) \rightarrow \mathfrak{u}_{j,2}(\mathbb{F}_{j,2})$, check if $\mathbb{x}$ satisfies the following condition: $(\mathbb{F}_{j,1} \in \mathbb{x}$ `and` $\mathbb{F}_{j,2} \in \mathbb{x})$ `or` $\mathbb{F}_{j,1} \notin \mathbb{x}$.

Step 2: Collect all the $\mathbb{x}$ that satisfy the above condition, and denote the collection as $\mathbb{D}_{\mathfrak{r}}$.

Step 3: The final selection dictionary is $\{\mathbb{A} \cup \mathbb{n} : \mathbb{n} \in \mathbb{D}_{\mathfrak{r}}\}$.

4.3.2 Variable selection via penalized regression

To conduct variable selection while respecting the defined selection rules, one can fit a penalized regression. However, how to define the grouping structure to satisfy a general selection rule is not well studied. To our knowledge, the most versatile penalized regression in terms of respecting selection rules is the latent overlapping group lasso (Obozinski et al., 2011b). We next show how to apply it to fulfill our objectives.

Denote the outcome by Y, the candidate covariate matrix by $\mathbf{X}$, and the model coefficients by $\boldsymbol{\beta}$. Note that the covariate matrix $\mathbf{X}$ corresponds to the full set of candidate variables, denoted by $\mathbb{V}$. Suppose we fit a logistic model $\text{logit}\{E(Y|\mathbf{X}; \beta_0, \boldsymbol{\beta})\} = \beta_0 + \mathbf{X}\boldsymbol{\beta}$. The grouping structure $\mathbb{G} \coloneqq \{\mathbb{g}_i, \cup_{i=1}^{I} \mathbb{g}_i = \mathbb{V}\}$ can be defined as a collection of non-empty subsets $\mathbb{g}_i \subseteq \mathbb{V}$ such that their union is $\mathbb{V}$. Define a set of latent variables $\bar{\boldsymbol{\alpha}} = (\boldsymbol{\alpha}^{\mathbb{g}_i})_{\mathbb{g}_i \in \mathbb{G}}$ such that $\sum_{i=1}^{I} \boldsymbol{\alpha}^{\mathbb{g}_i} = \boldsymbol{\beta}$, where $\boldsymbol{\alpha}^{\mathbb{g}_i}$ is a vector of the same length as $\boldsymbol{\beta}$ whose coordinates are non-zero for indices in the set $\mathbb{g}_i$ and 0 otherwise. The latent overlapping group lasso solves

$$\min_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}; D) + \sum_{\mathbb{g}_i \in \mathbb{G}} \omega_{\mathbb{g}_i} f(\boldsymbol{\alpha}^{\mathbb{g}_i}; \boldsymbol{\eta}),$$

where the level of penalization is controlled by the possibly multivariate hyper-parameter $\boldsymbol{\eta}$, $\omega_{\mathbb{g}_i}$ is a positive weight applied to the coefficients in the group $\mathbb{g}_i$, $\ell(\cdot)$ is a convex loss function (in our application, logistic loss (Cox, 1958)), and $f(\cdot)$ is a penalization function.

In practice, the convex loss function $\ell(\cdot)$ can be an L^2 norm with single hyperparameter $\boldsymbol{\eta} = \lambda$ (Obozinski et al., 2011b), minimax concave penalty (MCP) with hyperparameters $\boldsymbol{\eta} = (\lambda, \gamma)$ (Zhang, 2010; Huang et al., 2012), or smoothly clipped absolute deviation (SCAD) with $\boldsymbol{\eta} = (\lambda, \gamma)$ (Fan and Li, 2001; Breheny and Huang, 2015). The latter two penalties have the oracle property (Breheny and Huang, 2011) and retain the penalization rate of the L^2 norm for small coefficients, but continuously relax the rate of penalization as the absolute value of the coefficient increases. The rate of relaxation is larger in MCP, compared with SCAD.

The specifications of $f(\cdot)$ for these three penalties are given in supplementary material 1, Section B.2.

4.3.3 Constructing the grouping structure

In the latent overlapping group lasso, the selected variables must be the union of a subset of the groups in the grouping structure. (Obozinski et al., 2011b) Therefore, different specifications of grouping structures result in different combinations of subsets of variables that can be potentially selected. When the set of all combinations of group subsets is equivalent to the selection dictionary for the defined selection rules, we can say the grouping structure used with the latent overlapping group lasso respects the selection rules. (Wang et al., 2021) Correspondingly, Wang et al. (2021) developed a sufficient and necessary condition used with the latent overlapping group lasso for a grouping structure to respect a selection rule: the selection dictionary must be equal to $\{\cup_{\mathbb{g}\in\mathbb{Q}_j}\mathbb{g}, j = 1, \ldots, 2^I\}$, where $\mathbb{Q}_j, j = 1, ..., 2^I$ are all unique subsets of the grouping structure $\mathbb{G}$. However, the previous work does not show how to identify the grouping structure.

Based on Theorem 4 in Wang et al. (2021) and the nature of the latent overlapping group lasso, we develop roadmaps of grouping structure identification for some common selection rules, including the ones seen in this application. Define $\mathbb{A} = \{A_1, \ldots, A_n\}$, and $\mathbb{B} = \{B_1, \ldots, B_m\}$ as two non-overlapping sets of binary or continuous variables. Table 4.2 gives four types of selection rules with the corresponding roadmaps for constructing the corresponding grouping structure. More details of the roadmaps are given in Appendix B.3.

Since our selection rules 1.1-1.5 and 2 are represented in the form of "if one variable is selected, then some other variables must be selected", we will only need the first roadmap in Table 4.2. For example, for rule 1.4, we create single-variable groups for `DOAC` and `High-adherence` and a third group that contains `DOAC`, `High-adherence`, and their interaction. We do this for

Selection rule	**Roadmap of grouping variables**
If all variables in $\mathbb{A}$ are selected, then all variables in $\mathbb{B}$ must be selected	Specify m single-variable groups for $B_1 \ldots, B_m$, then specify another group that contains $\mathbb{A}$ and $\mathbb{B}$.
If at least one variable in $\mathbb{A}$ is selected, then all variables in $\mathbb{B}$ must be selected	Specify m single-variable groups for $B_1 \ldots, B_m$, then specify n groups, each containing $\mathbb{B}$ and A_i, for $i = 1, \ldots, n$.
If all variables in $\mathbb{A}$ are selected, then at least one variable in $\mathbb{B}$ must be selected	Specify m single-variable groups for $B_1 \ldots, B_m$, then specify m groups, each containing B_j and $\mathbb{A}$ for $j = 1, \ldots, m$.
If at least one variable in $\mathbb{A}$ is selected, then at least one variable in $\mathbb{B}$ must be selected	Specify m single-variable groups for $B_1 \ldots, B_m$, then specify $n \times m$ groups, where each group contains one variable from each set, i.e. each group contains A_i and B_j for some $i = 1, \ldots, n, j = 1, \ldots, m$.

Table 4.2: Roadmaps of grouping structure identification for the latent overlapping group lasso; $\mathbb{A} = \{A_1, \ldots, A_n\}$, and $\mathbb{B} = \{B_1, \ldots, B_m\}$ are two non-overlapping sets of binary or continuous variables.

each rule 1.1-2, removing duplicate groups. To implement rule 3, which cannot strictly be respected by latent overlapping group lasso, we modified it to "select either none or all of the 10 variables", which is practically the same as the original version (details in Appendix B.4). The modified rule can be implemented in the latent overlapping group lasso by specifying a group for those 10 variables.

Following the above steps, we established the grouping structure in the application, given in supplementary material 1, Section B.5. Then we used the `R` package `grpregOverlap` (Zeng and Breheny, 2016) to implement the latent overlapping group lasso with various penalties. The weights for each group were set to be the square root of the number of variables in the group. (Obozinski et al., 2011b) We next compare the results from the latent overlapping group lasso with those of lasso (Tibshirani, 1996) and adaptive lasso (Zou, 2006). Note that compared to lasso, adaptive lasso possesses the oracle property, namely, for large sample size, it performs as well as if the true underlying model were given in advance and thus

the results are more trustworthy. (Zou, 2006). All code for this analysis is available at https://github.com/Guanbo-W/SelectionDictionary.

4.4 Results

The rate of the outcome (major bleeding) is 3.47 per 100 person-years, and the percentage of patients who experienced major bleeding is 3.1%. Table 4.3 presents the means and standard errors or proportions of all variables stratified by the outcome. The summary statistics of variables stratified by high-dose-DOAC, low-dose-DOAC and warfarin are given in supplementary material 1 Section B.6.

Variable name (non-reference/reference level)	**Non-bleeding** n=20,671 (96.90%)	**Bleeding** n=661 (3.10%)
Baseline covariates		
1. Age, Mean (SD)	79.98 (8.79)	81.49 (7.93)
2. Sex (proportion female)	0.58	0.54
3. CHA_2DS_2-VASc Score ($\geqslant 3/<3$)	0.88	0.92
Comorbidities within 3 years before cohort entry		
4. Stroke(yes/no)	0.28	0.26
5. Anemia (yes/no)	0.10	0.17
6. Malignancy (yes/no)	0.26	0.30
7. Liver disease (yes/no)	0.02	0.05
8. History of major bleeding (yes/no)	0.34	0.49
9. Renal diseases (yes/no)	0.27	0.34
10. Heart disease(yes/no)	0.63	0.72
11. Diabetes (yes/no)	0.36	0.46
12. COPD/asthm (yes/no)	0.39	0.46
13. Dyslipidemia (yes/no)	0.58	0.64
OAC use at cohort entry		
14. DOAC (DOACs/warfarin)	0.57	0.50
15. Apixaban (yes/no)	0.30	0.21
16. Dabigatran (yes/no)	0.10	0.12

17. High-dose-DOAC (high-dose-DOACs/low-dose-DOACs or warfarin)	0.33	0.25
18. High-adherence (high/low)	0.87	0.25
19. Interaction of DOAC and High-adherence	0.49	0.43
20. Interaction of High-dose-DOAC and High-adherence	0.28	0.22
Concomitant medication use within 2 weeks before cohort entry		
21. Antiplatelets (yes/no)	0.32	0.40
22. NSAIDs (yes/no)	0.01	0.01
23. Antidepressants (yes/no)	0.18	0.22
24. PPIs (yes/no)	0.44	0.44
Potential drug-drug interaction*		
25. Interaction of DOAC and Antiplatelets	0.16	0.16
26. Interaction of DOAC and NSAIDs	0.01	0.01
27. Interaction of DOAC and Antidepressants	0.10	0.10
28. Interaction of DOAC and PPIs	0.23	0.19

*the variables are defined as the product of two drugs

Table 4.3: Covariate descriptive statistics (prevalences for binary covariates and means and standard errors for continuous covariates) stratified by the outcome

Crude (univariate analyses) and adjusted odds ratios (obtained from the logistic regression model adjusting for key covariates) and 95% confidence intervals are given in supplementary material 1, Section B.7.

Table 4.4 gives the odds ratios of each variable estimated by lasso, adaptive lasso, and latent overlapping group lasso, with the latter under different penalties. For all methods, we selected the tuning parameter λ at the minimum cross-validated risk, the tuning parameter γ of MCP and SCAD are set to 3 and 4 respectively.

	Non-grouped		**LOGL***		
Variable name	lasso	alasso**	L^2	MCP	SCAD
cross-validated risk	0.271	0.060	0.031	0.031	0.031
Baseline covariates					
1. Age ($\geqslant$75/$<$75)	1.21	1.09	1.24	1.24	1.24

2. Sex (female/male)	0.88	-	0.87	0.86	0.86
3. CHA_2DS_2-VASc Score ($\geqslant 3/<3$)	1.19	1.04	-	1.24	1.24
Comorbidities within 3 years before cohort entry					
4. Stroke (yes/no)	0.95	-	0.95	0.94	0.94
5. Anemia (yes/no)	1.35	1.35	1.35	1.38	1.38
6. Malignancy (yes/no)	1.07	-	1.08	1.09	1.09
7. Liver disease (yes/no)	1.92	1.98	1.92	2.01	2.01
8. History of major bleeding (yes/no)	1.59	1.60	1.53	1.62	1.62
9. Renal diseases (yes/no)	-	-	0.97	0.96	0.96
10. Heart disease (yes/no)	1.20	1.13	1.21	1.22	1.22
11. Diabetes (yes/no)	1.28	1.22	1.28	1.31	1.31
12. COPD/asthma (yes/no)	1.13	1.02	1.13	1.15	1.15
13. Dyslipidemia (yes/no)	1.07	-	1.07	1.09	1.09
OAC use at cohort entry					
14. DOAC (DOACs/warfarin)	1.18	-	1.29	1.39	1.39
15. Apixaban (yes/no), ref: Rivaroxaban	0.69	0.72	0.67	0.62	0.62
16. Dabigatran (yes/no), ref: Rivaroxaban	1.11	-	1.03	-	-
17. High-dose-DOAC (high-dose-DOACs/low-dose-DOACs or wafarin)	0.88	0.89	0.81	0.81	0.81
18. High-adherence (high-adherence/low-adherence)	-	-	1.01	-	-
19. Interaction of DOAC and High-adherence	-	-	0.96	-	-
20. Interaction of High-dose-DOAC and High-adherence	-	-	1.05	-	-
Concomitant medication use within 2 weeks before cohort entry					
21. Antiplatelets (yes/no)	1.35	1.17	1.30	1.51	1.51
22. NSAIDs (yes/no)	-	-	1.28	1.35	1.35
23. Antidepressants (yes/no)	1.13	-	1.15	1.16	1.16
24. PPIs (yes/no)	0.87	-	0.88	0.80	0.80
Potential drug-drug interaction					
25. Interaction of DOAC and Antiplatelets	0.81	-	0.86	0.68	0.68
26. Interaction of DOAC and NSAIDs	1.41	1.01	-	-	-
27. Interaction of DOAC and Antidepressants	-	-	0.94	-	-
28. Interaction of DOAC and PPIs	0.91	-	0.92	-	-

*the latent overlapping group lasso; ** adaptive lasso

Table 4.4: Coefficients estimates from various methods. - indicates the variable was not selected.

From the results, we see that lasso selected the most variables, while adaptive lasso selected the fewest variables. The resulting (averaged 10-fold) cross-validated risks were 0.271 and 0.060, respectively. However, neither method can incorporate selection rules. In fact, we see that both of them violated rule 3. In addition, both lasso and adaptive lasso selected the interaction of `DOAC` and `NSAIDs`, but not `NSAIDs`, violating strong heredity. Additionally, adaptive lasso selected `Dabigatran`, but not `DOAC`, which complicates the interpretation of the coefficient of `Dabigatran`. It selected `Dose` but not `DOAC`, which gives us a contrast between high-dose-DOACs versus low-dose-DOACs or warfarin, which is less interpretable than a contrast between high and low-dose-DOACs. Furthermore, adaptive lasso did not select `Sex`, `Stroke`, `Malignancy`, `Renal diseases`, which are well-known predictors of bleeding on oral anticoagulation, resulting in a model that would not be accepted by subject matter experts.

The latent overlapping group lasso reduced the cross-validated risk to 0.031. By design, the variable selection respects the selection rules, resulting in a model with better interpretability. Though the cross-validated risks of using different penalties in the latent overlapping group lasso were the same, the set of variables selected were different depending on the penalty type. The results derived from MCP and SCAD were similar, likely due to both of them having oracle properties. These methods penalize coefficients less when the estimated odds ratios deviate from 1. So we see that the estimated odds ratios from MCP and SCAD were further from 1 compared to those using the L^2 penalty. In addition, these two penalties resulted in fewer variables being selected. In contrast with the L^2 penalty, they did not select `High-adherence`, the interaction of `DOAC` and `High-adherence`, the interaction of `High-adherence` and `High-dose-DOAC`, the interaction of `DOAC` and `Antiplatelets`, and the interaction of `DOAC` and `PPIs`. Nevertheless, the estimated odds ratios of these variables

under the L^2 penalty were close to 1.

From MCP and SCAD, the most predictive factors associated with higher risk of major bleeding (odds ratios above 1.50) were `liver disease` (2.01), `History of major bleeding` (1.62) and `Antiplatelets` (1.51). Important predictors that were associated with lower risk of major bleeding (odds ratios below 0.8) are `Dabigatran` (0.62) and the interaction of `DOAC` and `Antiplatelets` (0.68). From the results, we can also summarize the estimated odds ratios of taking different types of DOACs versus warfarin in Table 4.5, which are also of interest. The method to obtain these additional contrasts is given in supplementary material 1, Section B.8.

Contrasts with warfarin as reference	Estimated odds ratios
High-dose-Apixaban	0.70
High-dose-Dabigatran	1.13
High-dose-Rivaroxaban	1.13
Low-dose-Apixaban	0.86
Low-dose-Dabigatran	1.39
Low-dose-Rivaroxaban	1.39

Table 4.5: Estimated odds ratios of taking different types of DOACs versus warfarin from the selected model by the latent overlapping group lasso MCP/SCAD

4.5 Discussion

This work represents the first application of the framework of structured variable selection developed by Wang et al. (2021), which we used to respect a complex series of selection rules while constructing a predictive model. Specifically, our application identified predictors of major bleeding among hospitalized hypertensive patients using OACs for atrial fibrillation using the data collected from administrative health databases.

We gave the implementation details for how to identify the selection dictionary related to the series of rules, and presented guidance in the form of roadmaps for how to develop an appropriate grouping structure for the latent overlapping group lasso. The `R` codes are also available, which provide a template to facilitate future applications. Overall, these tools can help to guide practitioners to carefully design selection rules informed by specific applications, derive the selection dictionary, identify the grouping structure and arrive at an interpretable predictive model.

The proposed methods can be easily implemented if the set of covariates that are affected by the selection rules is low-dimensional or can be divided into non-overlapping low-dimensional subsets, regardless of the dimension of the number of total variables. In such scenarios, one does not need to derive the selection dictionary for all variables. For example, suppose that only 10 variables are constrained by selection rules, and the other 1000 variables can be independently selected. We can first derive the sub-selection dictionary regarding these 10 variables (treating $\mathbb{V}$ as the set of these 10 variables), and then identify the sub-grouping structure for the 10 variables. Then the remainder of the complete grouping structure involves an additional 1000 single-variable groups for each of the remaining variables.

Even though the latent overlapping group lasso is the most versatile variable selection technique, there are some selection rules that it cannot respect. Examples include "select a number (between 0 and the cardinality of the subset) of variables in a subset of candidate variables", and "if a number of variables in a subset of candidate variables is selected, then select a number of variables in a (possibly distinct) subset of candidate variables". Given these limitations of the latent overlapping group lasso, future work could focus on the development of more general regularization methods that can respect an arbitrary selection rule. Another limitation of the application of the latent overlapping group lasso is that post-selection inference has not yet been developed for this method, so that post-selection confidence intervals are not currently available. Furthermore, the latent overlapping group lasso is also not able

to accommodate time-dependent covariates when the outcome is time-to-event. Future work involves applying the latent overlapping group lasso technique to the time-dependent Cox model.

In our development of an interpretable predictive model for major bleeding, we considered medication adherence, the interaction of adherence and DOAC, the dose of DOACs, and the interaction between the dose of DOACs and adherence. In addition, we also incorporated selection dependencies arising from the selection of drug-drug interactions. For all interactions, we applied strong heredity, resulting in a complex set of selection rules that our method was able to respect. In contrast, the selected models resulting from standard lasso and adaptive lasso violated selection rules, and thus lacked interpretability.

Interpretable prediction modeling can guide clinicians in identifying patients at risk of an outcome by highlighting which factors are predictive of the outcome of interest and their importance in prediction. However, when the unique goal of modeling is the determination of patient risk, black box methods (i.e. those not restricted to a semiparametric model to the extent that the fitted model cannot be interpreted by the ultimate user) may be more powerful. They may thus be preferred as a decision-making aid. However, much clinical practice remains unassisted by such algorithms and thus interpretable modeling continues to provide important contributions to medical knowledge.

Supplementary materials

Supplementary materials 1 contains some technical details of the method and numerical results from the analysis.

Supplementary materials 2 is the variable definitions of variables considered in the analysis.

Codes for deriving the selection dictionary, and analysis, and the resulting selection dictio-

nary are seen in the Github https://github.com/Guanbo-W/SelectionDictionary.

Data availability statement

The data analyzed in this work are not allowed to share due to the confidentiality.

Funding

The study was supported by the Heart and Stroke Foundation of Canada (G-17-0018326) and the Réseau Québécois de Recherche sur les Médicaments (RQRM). Please refer to https://www.heartandstroke.ca/ and http://www.frqs.gouv.qc.ca/en/. G Wang is supported by a Fonds de Recherche du Québec—Santé Doctoral Training Award (272161) and the Faculté de Pharmacie, Université de Montréal. ME Schnitzer is supported by a Canadian Institutes of Health Research Canada Research Chair tier 2 and a Natural Sciences and Engineering Research Council of Canada Discovery Grant. RW Platt is supported by a CIHR Foundation Grant (FDN-143297).

Acknowledgements

We would like to thank the RAMQ and Quebec Health Ministry for providing assistance in handling the data and the Commission d'accès à l'information for authorizing the study.

Chapter 5

Structured variable selection in Cox model with time-dependent covariates

This chapter is not a stand-alone manuscript, much of the literature review can be found in Chapter 2. The overlapping group Lasso was developed and applied mainly in the context of engineering and machine learning, where the outcome is often continuous or binary. In medical research, one common outcome of interest is time-to-event. Therefore, to broaden the use of the overlapping group Lasso, in this chapter, we extend it to accommodate data consisting of survival outcomes and time-dependent covariates. We present the algorithm of the overlapping group Lasso in the context of survival data in a straightforward way, which avoids knowledge of graph models. In the simulation study, we design complex covariate structures and show how to use the group variables based on Theorem 5 in Chapter 3. Various metrics of the performance of the estimator are assessed, and compared to the standard Lasso penalization in time-dependent Cox.

Note that the supplementary material for this chapter can be found in Appendix C.

5.1 Methods

In this section, we present the specific statistical objective and demonstrate how to solve the problem by leveraging the algorithm developed by Mairal et al. (2010) while incorporating backtracking line search, and connect the problem to the framework developed in Chapter 3.

5.1.1 The objective function

We reviewed the overlapping group Lasso and the time-dependent Cox model in Sections 2.2.8, and 2.3.3, respectively. To accommodate the time-dependent Cox model into the overlapping group Lasso, we essentially need to solve the following problem

$$\min_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) + \lambda \Omega(\boldsymbol{\beta}), \tag{5.1}$$

where

$$f(\boldsymbol{\beta}) = -\frac{1}{n}\log\{L(\boldsymbol{\beta})\} = -\sum_{j=1}^{m}\left(\left\{\sum_{l\in D_j}\boldsymbol{X}_l(t_j)\right\}\boldsymbol{\beta} - d_j \log\left[\sum_{l\in R_j}\exp\{\boldsymbol{X}_l(t_j)\boldsymbol{\beta}\}\right]\right)$$

is the averaged negative log partial likelihood defined in (2.9), and $\Omega(\boldsymbol{\beta}) = \sum_{\mathbb{g}\in\mathbb{G}}\omega_{\mathbb{g}}\left\|\boldsymbol{\beta}_{|\mathbb{g}}\right\|_{\infty}$ is the weighted sum of L^{∞} norms of pre-defined groups of coefficients $\boldsymbol{\beta}_{|\mathbb{g}}, \forall \mathbb{g} \in \mathbb{G}$. While $f(\boldsymbol{\beta})$ is a convex differentiable function, $\Omega(\boldsymbol{\beta})$ is not differentiable on all of its support. Thus, the optimization of the penalized likelihood requires the proximal method.

The main difference between our proposed method and the overlapping group Lasso is the definition of $f(\boldsymbol{\beta})$. Since we consider survival data, $f(\boldsymbol{\beta})$ is not a square or logistic loss, which results in the difference in estimating the model coefficients.

5.1.2 Optimization

To solve the above optimization problem, we need to overcome two challenges: 1) a non-differentiable penalty and 2) a penalty consisting of overlapping groups of coefficients. We next give the details of how to leverage a proximal operator (and its duality), network flow algorithms, and backtracking line search to solve the problem in our context. For accessibility for a statistics readership, we sketch the network flow algorithm in a relatively straightforward way avoiding knowledge of theorems in graph models.

The proximal method and its duality

The proximal method (Moreau, 1962) has been successfully applied in many research areas including signal processing (Wright et al., 2009; Becker et al., 2011; Combettes and Pesquet, 2011) and machine learning (Bach, 2010). In order to conquer the computational problem caused by the non-smooth component in the objective function, in each iteration, instead of updating the estimate with respect to the gradient, it uses a *proximal operator*, so that the updated estimates stay close to the gradient update for the differentiable function, while also making the non-differentiable function small. (Beck and Teboulle, 2009; Nesterov, 2013a) It is proven to converge at fairly fast rates. (Nesterov, 2007; Beck and Teboulle, 2009; Beck, 2017)

The proximal operator *Prox* is defined as

$$Prox_{t,\lambda,\Omega}(\boldsymbol{u}) = \underset{\boldsymbol{v}}{\operatorname{argmin}} \frac{1}{2t} \|\boldsymbol{u} - \boldsymbol{v}\|_2^2 + \lambda\Omega(\boldsymbol{v}).$$

We then approach the optimization problem (5.1) by proximal gradient descent. In each iteration of the following algorithm, let the updated value of $\boldsymbol{\beta}$ be $\boldsymbol{\beta}^+$ and the current value

of $\boldsymbol{\beta}$ be $\tilde{\boldsymbol{\beta}}$. Then using a Taylor expansion (Moreau, 1962), we have

$$\begin{aligned}\boldsymbol{\beta}^{+} =& \underset{\boldsymbol{\beta}}{\operatorname{argmin}} f(\tilde{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})\nabla f(\tilde{\boldsymbol{\beta}}) + \lambda\Omega(\boldsymbol{\beta}) + \frac{1}{2t}\left\|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\right\|_2^2 \\ =& \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{2t}\left\|\boldsymbol{\beta} - \{\tilde{\boldsymbol{\beta}} - t\nabla f(\tilde{\boldsymbol{\beta}})\}\right\|_2^2 + \lambda\Omega(\boldsymbol{\beta}) \\ =& Prox_{t,\lambda,\Omega}\left\{\tilde{\boldsymbol{\beta}} - t\nabla f(\tilde{\boldsymbol{\beta}})\right\},\end{aligned} \tag{5.2}$$

where t is a step size parameter and the upper bound of $\frac{1}{t} > 0$ is the Lipschitz constant of $\nabla f(\boldsymbol{\beta})$.

In many cases, the proximal operator has closed form solutions, so it can be implemented with ease. However, with an overlapping grouping structure where some coefficients are included in different groups of penalties, the closed form solution does not exist (Jenatton et al., 2011b). Alternatively, the proximal operator can be expressed as the residual of the projection of a vector onto a ball of the dual-norm $\|\cdot\|_*$, where $\|\kappa\|_* \coloneqq \max_{\|\boldsymbol{z}\|\leqslant 1}\boldsymbol{z}^T\kappa$. (Wright et al., 2009; Jenatton et al., 2011b; Combettes and Pesquet, 2011) Closely following Jenatton et al. (2011b), we next derive the dual of (5.2) in Lemma 1, and provide the conditions under which the primary-dual variables are optimal. The proof and additional details on Lemma 1 are provided in the Appendix C.1.

LEMMA 1. *The right hand side of (5.2) can be expressed by a dual formulation*

$$\underset{\xi\in\mathbb{R}^{p\times|\mathbb{G}|}}{\operatorname{argmin}} \frac{1}{2t}\left[\left\|\left\{\tilde{\boldsymbol{\beta}} - t\nabla f(\tilde{\boldsymbol{\beta}})\right\} - \sum_{\mathfrak{g}\in\mathbb{G}}\boldsymbol{\xi}_{|\mathfrak{g}}\right\|_2^2 - \left\|\tilde{\boldsymbol{\beta}} - t\nabla f(\tilde{\boldsymbol{\beta}})\right\|_2^2\right] \quad s.t.\forall\mathfrak{g}\in\mathbb{G}, \left\|\boldsymbol{\xi}_{|\mathfrak{g}}\right\|_* \leqslant \lambda\omega_{\mathfrak{g}} \tag{5.3}$$

with dual variable $\boldsymbol{\xi}$, meaning that strong duality holds between (5.2) and (5.3). The pair of

primal-dual variables $\{\boldsymbol{\beta}, \boldsymbol{\xi}\}$ *is optimal if and only if* $\boldsymbol{\xi}$ *satisfies the constraint in (5.3), and*

$$\boldsymbol{\beta} = \left\{\tilde{\boldsymbol{\beta}} - t\nabla f(\tilde{\boldsymbol{\beta}})\right\} - \sum_{\mathbb{g}\in\mathbb{G}} \boldsymbol{\xi}_{|\mathbb{g}}, \forall \mathbb{g} \in \mathbb{G}, \tag{5.4}$$

$$\boldsymbol{\xi}_{|\mathbb{g}} = \prod_{\|\cdot\|_*\leqslant\lambda\omega_{\mathbb{g}}} (\boldsymbol{\beta}_{|\mathbb{g}} + \boldsymbol{\xi}_{|\mathbb{g}}) = \prod_{\|\cdot\|_*\leqslant\lambda\omega_{\mathbb{g}}} \left(\left[\tilde{\boldsymbol{\beta}} - t\nabla f(\tilde{\boldsymbol{\beta}}) - \sum_{\mathbb{h}\neq\mathbb{g}} \boldsymbol{\xi}_{|\mathbb{h}}\right]_{|\mathbb{g}} \right), \tag{5.5}$$

where $\prod_{\|\cdot\|_*\leqslant\lambda\omega_g}$ *denotes the orthogonal projection onto the ball* $\{\boldsymbol{\kappa} \in \mathbb{R}^p; \|\boldsymbol{\kappa}\|_* \leqslant \lambda\omega_g\}$.

While the Lemma makes it possible to solve the optimization problem, how to compute $\sum_{\mathbb{g}\in\mathbb{G}} \boldsymbol{\xi}_{|\mathbb{g}}$ is not trivial. This is because of the overlapping structure among the $\boldsymbol{\beta}_{|\mathbb{g}}$s, and thus the $\boldsymbol{\xi}_{|\mathbb{g}}$s. Next we show how the network flow algorithm can overcome this challenge.

Network flow algorithm

Different approaches have been proposed to solve the proximal operator when the groups are in special structures. For instance, when all groups are nested (called a *tree structure*), the dual can be computed by block coordinate ascent. (Bertsekas et al., 1999; Jenatton et al., 2011b) However, the general case of overlapping groups is more challenging.

Mairal et al. (2010) tackled the problem by a network flow algorithm. The authors developed and evaluated the method based on graph models, and applied it to image processing problems. However, the relevant technical development is rather involved and thus hinders understanding for readers without sufficient background in computer science and operational research. Though the target question can be well formulated by graph models, it does not have to be in our scenario, and thus we present it in a way that circumvents understanding of the complex theorems.

Briefly, based on Lemma 1, Mairal et al. (2010) showed that the target problem is dual to a quadratic min-cost flow problem, which can be solved by the mini-cut theorem (Ford and Fulkerson, 1956). For readers who enjoy technical details, additional relevant theorems can

be found in (Goldberg and Tarjan, 1988; Cherkassky and Goldberg, 1997; Bertsekas, 1998; Mairal et al., 2010).

Based on Lemma 1 and the network flow algorithm, we next provide the algorithm adapted to our context, to solve (5.2) in Algorithm 3. Informally, the network flow algorithm obtains $\sum_{\mathbb{g}\in\mathbb{G}} \boldsymbol{\xi}_{|\mathbb{g}}$ by applying the function `computeFlow`, which is based on the following steps:

1. Projection step: find the projection of the vectors $\boldsymbol{\xi}_{|\mathbb{g}}$, $\boldsymbol{\gamma} = (\gamma_1, \dots)$. It is done by solving a relaxed version of (5.3), which finds the value of $\boldsymbol{\gamma}$ which is the lower bound of $\frac{1}{2t}\left\|\left\{\tilde{\boldsymbol{\beta}} - t\nabla f(\tilde{\boldsymbol{\beta}})\right\} - \boldsymbol{\gamma}\right\|_2^2$, and $\sum_j \gamma_j \leqslant \lambda \sum_{|\mathbb{g}\in\mathbb{G}} \omega_{|\mathbb{g}}$.

2. Updating step: update $(\sum_{\mathbb{g}\in\mathbb{G}} \boldsymbol{\xi}^j_{|\mathbb{g}})_{\boldsymbol{x}_j\in\mathbb{V}}$ by maximizing $\sum_{\boldsymbol{x}_j\in\mathbb{V}} \sum_{\mathbb{g}\in\mathbb{G}} \boldsymbol{\xi}^j_{|\mathbb{g}}$, while keeping $\sum_{\boldsymbol{x}_j\in\mathbb{g}} \boldsymbol{\xi}^j_{|\mathbb{g}} \leqslant \lambda\omega_{\mathbb{g}}$. By doing so, we can ensure that the constraint in (5.3) holds. This can be done by the max flow algorithm. Details of the implementation can be found in Appendix C.2.

3. Recursion step/divide and conquer: According to the mini-cut theorems (Ford and Fulkerson, 1956), define $\mathbb{V}^* = \{\boldsymbol{x}_j \in \mathbb{V} : \sum_{\mathbb{g}\in\mathbb{G}} \boldsymbol{\xi}^j_{|\mathbb{g}} = \gamma_j\}$, and $\mathbb{G}^* = \{\mathbb{g} \in \mathbb{G} : \sum_{\boldsymbol{x}_j\in\mathbb{g}} \boldsymbol{\xi}^j_{|\mathbb{g}} < \lambda\omega_{\mathbb{g}}\}$. Then apply steps 1 and 2 to $(\mathbb{V}^*, \mathbb{G}^*)$ and their respective complements until $(\sum_{\mathbb{g}\in\mathbb{G}} \boldsymbol{\xi}^j_{|\mathbb{g}})_{\boldsymbol{x}_j\in\mathbb{V}}$ (obtained from step 2) matches $\boldsymbol{\gamma}$ (obtained from step 1).

The algorithm is equivalent to the network flow algorithm, but with a different presentation. Therefore, the algorithm enjoys the fast convergence property as well. However, the presentation of the algorithm 1) connects to the general framework for structured variable selection, and is consistent with the research problem in our context, 2) offers the possibility of using backtracking line search and 3) is helpful to understand the skeleton as well as the essence of the network flow algorithm.

Algorithm 3 Solving (5.2) using quadratic min-cox flow

Inputs: The estimate in the kth step $\boldsymbol{\beta}^k \in \mathbb{R}^p$, step size t, set of variables $\mathbb{V}$ and groups $\mathbb{G}$, group weights $\omega_{\mathbb{g}}$, regularization parameter λ

Set $\boldsymbol{\xi} = 0$

Compute $\sum_{\mathbb{g}\in\mathcal{G}} \boldsymbol{\xi}_{|\mathbb{g}} \leftarrow \texttt{computeFlow}(\mathbb{V}, \mathbb{G})$

return $\boldsymbol{\beta}^k - t\nabla f(\boldsymbol{\beta}^k) - \sum_{\mathbb{g}\in\mathcal{G}} \boldsymbol{\xi}_{|\mathbb{g}}$

Function `computeFlow`$(\mathbb{V}, \mathbb{G})$
Projection step: $\boldsymbol{\gamma} \leftarrow \text{argmin}_{\boldsymbol{\gamma}} \sum_{j:\boldsymbol{x}_j\in\mathbb{V}} \frac{1}{2t}(\beta_j^k - t\nabla f(\beta_j^k) - \gamma_j)^2$ s.t. $\sum_{j:\boldsymbol{x}_j\in\mathbb{V}} \gamma_j \leqslant \lambda \sum_{\mathbb{g}\in\mathbb{G}} \omega_{\mathbb{g}}$

Updating step: update $(\sum_{\mathbb{g}\in\mathbb{G}} \boldsymbol{\xi}^j_{|\mathbb{g}})_{\boldsymbol{x}_j\in\mathbb{V}} \leftarrow \text{argmax}_{(\sum_{\mathbb{g}\in\mathbb{G}} \boldsymbol{\xi}^j_{|\mathbb{g}})_{\boldsymbol{x}_j\in\mathbb{V}}} \sum_{\boldsymbol{x}_j\in\mathbb{V}} \sum_{\mathbb{g}\in\mathbb{G}} \boldsymbol{\xi}^j_{|\mathbb{g}}$
s.t. $\sum_{\boldsymbol{x}_j\in\mathbb{g}} \boldsymbol{\xi}^j_{|\mathbb{g}} \leqslant \lambda\omega_{\mathbb{g}}$

Recursion step:
if $\exists \boldsymbol{x}_j \in \mathbb{V}$ *s.t.* $, \sum_{\mathbb{g}\in\mathbb{G}} \boldsymbol{\xi}^j_{|\mathbb{g}} \neq \gamma_j$ **then**
  Denote $\mathbb{V}^* = \{\boldsymbol{x}_j \in \mathbb{V} : \sum_{\mathbb{g}\in\mathbb{G}} \boldsymbol{\xi}^j_{|\mathbb{g}} = \gamma_j\}$, and $\mathbb{G}^* = \{\mathbb{g} \in \mathbb{G} : \sum_{\boldsymbol{x}_j\in\mathbb{g}} \boldsymbol{\xi}^j_{|\mathbb{g}} < \lambda\omega_{\mathbb{g}}\}$
  $(\sum_{\mathbb{g}\in\mathbb{G}} \boldsymbol{\xi}^j_{|\mathbb{g}})_{\boldsymbol{x}_j\in\mathbb{V}^*} \leftarrow \texttt{computeFlow}(\mathbb{V}^*, \mathbb{G}^*)$
  $(\sum_{\mathbb{g}\in\mathbb{G}} \boldsymbol{\xi}^j_{|\mathbb{g}})_{\boldsymbol{x}_j\in\mathbb{V}\setminus\mathbb{V}^*} \leftarrow \texttt{computeFlow}(\mathbb{V}\setminus\mathbb{V}^*, \mathbb{G}\setminus\mathbb{G}^*)$
end
return $(\sum_{\boldsymbol{x}_j\in\mathbb{g}} \boldsymbol{\xi}^j_{|\mathbb{g}})_{\boldsymbol{x}_j\in\mathbb{V}}$

The bold Greek letters denote vectors of length p, and ones with subscript j (or superscript for $\boldsymbol{\xi}_{|\mathbb{g}}$) denotes the value of the vector in the correspond indexed position.

Backtracking line search

Having the algorithm to solve the dual of the proximal operator, we can thus use proximal gradient descent with backtracking line search (Bertsekas, 1997) to solve (5.1). Backtracking line search is a technique in optimization that can determine the step size properly. It starts with a pre-defined step size for updating along the search direction, and shrinks the step size (i.e., "backtracking") iteratively until a decrease of the loss function fairly corresponds to the expected decrease, based on the local gradient of the loss function. Moreover, it speeds up

the convergence.

Different from the work by Mairal et al. (2010) and Jenatton et al. (2011b), the proximal operator defined in (5.2) includes the step size t, which makes it possible for us to incorporate backtracking line search. The proposed algorithm is shown in Algorithm 4.

Algorithm 4 Solving (5.1) using proximal gradient descent with backtracking line search

Inputs: covariate $\boldsymbol{x}_i(t)(i = 1, \ldots, n)$, survival time T_i, censored indicator δ_i, set of variables $\mathbb{V}$ and groups $\mathbb{G}$, group weights $\omega_{\mathbb{g}}$, regularization parameter λ, convergence threshold r, shrinkage rate $\alpha < 1$, step size t

Set $\boldsymbol{\beta}^0 = \mathbf{0}, k = 0$

repeat
  $\boldsymbol{\beta}^+ \leftarrow Prox_{t,\lambda,\Omega}\left(\boldsymbol{\beta}^k - t\nabla f(\boldsymbol{\beta}^k)\right)$ ▷ call Algorithm 3
  if $f(\boldsymbol{\beta}^+) \leqslant f(\boldsymbol{\beta}^k) + \nabla f(\boldsymbol{\beta}^k)^\intercal(\boldsymbol{\beta}^+ - \boldsymbol{\beta}^k) + \frac{1}{2t}\left\|\boldsymbol{\beta}^+ - \boldsymbol{\beta}^k\right\|_2^2$
  then
    $k = k + 1$;
    $\boldsymbol{\beta}^{k+1} \leftarrow \boldsymbol{\beta}^+$
    exit;
  else
    $t \leftarrow \alpha t$
  end
until $\left\|\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1}\right\|_1 < r$;
return $\hat{\boldsymbol{\beta}} \leftarrow \boldsymbol{\beta}^{k+1}$

5.2 Simulation

Our first goal in this simulation is to apply the framework in Chapter 3 to derive the desired selection dictionary and grouping structure to incorporate the needed selection rules with a relatively complex data structure. Another goal of the simulation study is to empirically evaluate our method's properties. This is done by comparing our penalty with the included grouping structure to the unstructured L^1 penalty when implemented in a Cox model with time-dependent covariates. In this simulation, we generate data which contain three main

terms (two of them are categorical variables) and two interactions (also of categorical variables).

5.2.1 Simulation design

Covariate

We generate 3 independent variables (potential predictors) A, B and C, where the values of each variable for each subject randomly change over time in a piece-wise constant fashion. We include 50 time points at which the values of any variable can potentially change, but we hold the value of a variable constant for at least 5 and at most 10 time points. A and C are three-level categorical variables, represented by two dummy variables denoted by A_1 and A_2, and C_1 and C_2, respectively. B is a continuous variable.

See below the algorithm for generating A and C. B was generated in a similar fashion except that in Step 1, we generated a series of numbers following a standard normal distribution.

Algorithm 5 Steps for generating each time-dependent categorical variable

For each subject i:

Step 1: with replacement, sample 10 integers from 1, 2, 3 with equal probability to represent the three categories of the variables.

Step 2: for each sampled value, repeat the value R_i times, where R_i is sampled from 5, 6, ..., 10 with equal probabilities. Then, concatenate these repeated values all together, resulting in a single vector with length between 50 and 100.

Step 3: take the first 50 elements as the values of the categorical variable.

Outcome

We generated the time-to-event outcome by a permutation algorithm developed by Sylvestre and Abrahamowicz (2008) using the `R` function `PermAlgo` (Sylvestre et al., 2010). The generated event times are dependent on time-dependent potential predictors according to

the proportional hazards model

$$h(t|\boldsymbol{X}) = h_0(t)\exp\{\boldsymbol{\beta}^\intercal \boldsymbol{x}(t)\},$$

where $h(t|\boldsymbol{X})$ and $h_0(t)$ are the hazard and baseline hazard at time t, $\boldsymbol{x}(t)$ is the vector of predictor values at time t, $\boldsymbol{X} = \{\boldsymbol{x}(t), t = 1, \ldots .50\}$, and $\boldsymbol{\beta}$ is the vector of log hazard ratios of the predictors. The outcome had around 50% random censoring. The median event time was at around time 25 among those not censored.

The (potential) predictors in the Cox model were A, B, C, and two interactions 1) AB: the interaction of A and B, and 2) BC: the interaction of B and C. Recall that all covariates except B are three-level categorical variables. We thus have $\boldsymbol{x}(t) = (A_1, A_2, B, A_1B, A_1B, C_1$, C_2 , C_1B , $C_2B)$, and $\boldsymbol{\beta}$ is a vector of length 9.

We generated two scenarios regarding the $\boldsymbol{\beta}$:

1. $\boldsymbol{\beta}$ =(log(3), log(3), 0, 0, 0, 0, 0, 0, 0). That is, only A_1 and A_2 are predictive of the outcome, which corresponds to a sparse covariate structure.

2. $\boldsymbol{\beta}$ =(log(3), log(3), log(4), log(3), log(3), 0, 0, 0, 0). That is, there are five true predictors $(A_1, A_2, B, A_1B, A_2B)$ and four noise variables, which corresponds to a less sparse covariate structure.

Selection rules and grouping structure

In our method, we incorporated the selection rules:

1. The dummy variables representing a categorical variable should be selected collectively.

2. If an interaction is selected, the main terms must be selected, and

To respect the selection rules, we followed Chapter 3 and derived the selection dictionary using `R` by following Table 3.2 in Chapter 3. Our resulting selection dictionary is: $\{\emptyset$, $\{B\}$,

$\{C_1, C_2\}$, $\{B, C_1, C_2\}$, $\{B, C_1, C_2, C_1B, C_2B\}$, $\{A_1, A_2\}$, $\{A_1B, A_2B, B\}$, $\{A_1, A_2, C_1$, $C_2\}$, $\{A_1, A_2, A_1B, A_2B, B\}$, $\{A_1, A_2, B, C_1, C_2\}$, $\{A_1, A_2, B, C_1, C_2, A_1B, A_2B\}$, $\{A_1, A_2, B$, $C_1, C_2, C_1B, C_2B\}$, $\{A_1, A_2$, $B, A_1B, A_2B, C_1, C_2, C_1B, C_2B\}\}$. The `R` code is available at the GitHub repository linked below. In Appendix C.3, we show how to use the selection dictionary to determine the grouping structure. The result is given below.

$$\mathcal{G}_1 = \{A_1, A_2, A_1B, A_2B\}, \mathcal{G}_2 = \{B, A_1B, A_2B, C_1B, C_2B\},$$
$$\mathcal{G}_3 = \{A_1B, A_2B\}, \mathcal{G}_4 = \{C_1, C_2, C_1B, C_2B\}, \mathcal{G}_5 = \{C_1B, C_2B\}.$$

One can verify the correctness of the derived selection dictionary using Theorem 5 in Chapter 3.

Sample size

We included four settings, with sample sizes N=100, 500, 1000, and 2000 respectively.

Comparison

The `R` package `glmnet` (Friedman et al., 2010a; Simon et al., 2011) can perform "Lasso" (by penalizing by the L^1 norm of coefficients) in Cox models with time-dependent covariates. We thus compared this method with ours. That is, we compared our method with unstructured variable selection. We also compared two methods to select the tuning parameter: the "one-standard-error-rule" (1se) and the "select the tuning parameter which has the minimum averaged cross-validated error" (min). We denote our method used with 1se and min by OLG.1se and OLG.min respectively, and Lasso used with 1se and min by L.1se and L.min respectively.

Performance statistics

We used the following measures to compare the performance of the two methods. Each measure is calculated for each individual simulated dataset, then averaged to obtain the performance statistics.

1. Joint Detection Rate (JDR): binary, JDR=1 if all the selected variables are true predictors (but also possibly selected some other noise variables), 0 otherwise.
2. Missing Rate (MR): the percentage of variables not selected among the true predictors.
3. False Alarm Rate (FAR): the percentage of selected variables among the noise variables.
4. Consistency of Categorical Variable Selection (CCVS): indicator of whether the resulting selected model satisfied selection rule 1.
5. Consistency of Strong Hierarchy (CSH): indicator of whether the resulting selected model satisfied selection rule 2.
6. Refitted C Index (RCI): The C index of the model with the selected variables.
7. Mean Squared Errors (MSE): mean squared difference between the coefficients in the data generating mechanism and the estimates resulting from implementing the method.
8. Averaged cross-validated errors (CV-E): cross-validated error.

The CV-E is defined as follows. Suppose we perform K-fold cross-validation, denote $\hat{\beta}^{-k}$ by the estimate obtained from the rest of K-1 folds (training set). The error of the k-th fold (test set) is defined as 2(P-Q)/R, where P is the log partial likelihood evaluated at $\hat{\beta}^{-k}$ using the entire dataset, Q is the log partial likelihood evaluated at $\hat{\beta}^{-k}$ using the training set, and R is the number of events in the test set. We do not use the negative log partial likelihood evaluated at $\hat{\beta}^{-k}$ using the test set because the former definition can efficiently use the risk set, and thus it is more stable when the number of events in each test set is small (think of leave-one-out). The CV-E is used in parameter tuning. To account for balance in outcomes

among the randomly formed test set, we divide the deviance 2(P-Q) by R.

Each statistic in Table 5.1 represents the mean of the above measurements among 100 runs.

Implementation Details

The simulation was performed in `R` version 4.0.5 (R Core Team, 2021) and the code is available at `https://github.com/Guanbo-W/StructuralTDCox`. The code calls the `R` functions `spams.proximalGraph` (for C++ SPAMS code et al., 2017) and `coxph` (Therneau, 2021).

All simulations used 10-fold cross-validation. The weight ω_g for each group in the penalization term was set equal to one. In the cross-validation, we needed to specify the range of the tuning parameter λ. In scenario 1, λ values ranged from 0.0001 to 0.2, and in scenario 2 from 0.0001 to 0.5. The λ sequences increase on a log scale. In the implementation of our algorithm in the simulation, the step size and step size shrinkage rate were set to 1 and 0.8, respectively. The convergence criterion (the sum of absolute difference between the estimates from the two steps) was 10^{-5}. Each simulation had 100 runs (double the number used in Tibshirani 1997).

5.2.2 Simulation Results

The results are given in Table 5.1. Our method's CCVS and CSH were always 1, which means the results from our method always respected the given selection rules. This was not true when using the unstructured L^1 penalty. L.min failed to respect the categorical selection rule 20%-40% of the time, while it broke the strong heredity rule 80% of the time in some settings; this did not improve with increased sample size. Though the performance of L.1se was better, and though it may respect rules with a very large sample size when the rules are satisfied in the true model, it cannot guarantee it in finite samples. Note that

Scenario	**1 (A≠0)**				**2 (A, B, AB≠0)**			
Method	OLG.1se	OLG.min	L.1se	L.min	OLG.1se	OLG.min	L.1se	L.min
	N=100							
JDR	0.18	0.98	0.00	0.66	0.90	1.00	0.02	0.44
MR	0.82	0.02	0.85	0.26	0.08	0.00	0.41	0.16
FAR	0.02	0.65	0.03	0.40	0.04	0.53	0.06	0.49
CCVS	1.00	1.00	0.88	0.70	1.00	1.00	0.84	0.72
CSH	1.00	1.00	0.75	0.38	1.00	1.00	0.39	0.35
RCI	0.53	0.67	0.54	0.64	0.91	0.92	0.91	0.92
MSE	0.24	0.06	0.27	0.14	0.27	0.10	0.41	0.30
CV-E	6.87	6.74	6.68	6.59	4.70	4.36	4.74	4.33
	N=500							
JDR	0.88	1.00	0.50	1.00	1.00	1.00	0.08	0.98
MR	0.12	0.00	0.32	0.00	0.00	0.00	0.36	0.04
FAR	0.02	0.70	0.01	0.54	0.00	0.15	0.05	0.59
CCVS	1.00	1.00	0.90	0.57	1.00	1.00	0.94	0.78
CSH	1.00	1.00	0.79	0.22	1.00	1.00	0.45	0.62
RCI	0.60	0.63	0.58	0.63	0.91	0.91	0.91	0.91
MSE	0.14	0.02	0.22	0.03	0.14	0.04	0.31	0.07
CV-E	6.79	6.70	6.81	6.68	4.38	4.26	4.40	4.26
	N=1000							
JDR	1.00.	1.00.	0.94	1.00	1.00	1.00	0.20	0.98
MR	0.00	0.00	0.04	0.00	0.00	0.00	0.20	0.02
FAR	0.02	0.73	0.02	0.58	0.00	0.08	0.04	0.66
CCVS	1.00	1.00	0.97	0.70	1.00	1.00	0.90	0.82
CSH	1.00	1.00	0.94	0.31	1.00	1.00	0.53	0.65
RCI	0.61	0.62	0.60	0.62	0.91	0.91	0.91	0.91
MSE	0.10	0.01	0.15	0.02	0.12	0.02	0.29	0.08
CV-E	6.76	6.68	6.76	6.67	4.33	4.25	4.36	4.24
	N=2000							
JDR	1.00	1.00	1.00	1.00	1.00	1.00	0.74.	1.00
MR	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00
FAR	0.00	0.06	0.00	0.54	0.00	0.00	0.04	0.57
CCVS	1.00	1.00	1.00	0.60	1.00	1.00	0.93	0.76
CSH	1.00	1.00	0.99	0.20	1.00	1.00	0.79	0.57
RCI	0.61	0.61	0.61	0.61	0.91	0.91	0.91	0.91
MSE	0.06	0.00	0.10	0.01	0.12	0.05	0.23	0.01
CV-E	6.71	6.66	6.72	6.66	4.30.	4.23	4.31.	4.23

L: non-structured L^1 penalty (`glmnet` with cox); OLG: our method; 1se: applying "one-standard-error-rule"; min: selecting the model with the least CV-E.

Table 5.1: Simulation results

in scenario 2 (where the true data generating mechanism involved the coefficients of one interaction and its main terms all being non-zero), the unstructured L^1 penalty had a higher chance of breaking the rules as compared to scenario 1.

If applying the min rule, which selects the model that has the lowest CV-E, the JDR and MR in both methods can achieve 1 and 0 respectively relatively fast with increasing sample sizes, though ours was faster. However, if applying the 1se rule, while the JDR and MR of our method converged to 1 and 0 respectively relatively fast, this was not the case under the unstructured L^1 penalty. This means that by using our method, we have a higher chance of successfully selecting the variables that should be selected (even with a relatively small sample size).

For both methods, the FAR deviated from 0 when not using the 1se rule, meaning that the methods selected many variables whose coefficients are actually zero in the data generating mechanism. This is because the CV-E can be regarded as the "prediction accuracy". Our setting is not a high-dimensional setting, so if the goal is to recover the sparsity pattern, one should avoid selecting the model that has the lowest CV-E (though the FAR of our method converged to 0 relatively fast with the increasing sample size). If applying the 1se rule, we observed that in the less sparse setting (scenario 2), our method had a lower chance of selecting the variables that should not be selected.

The simulation results also showed that our method can achieve better estimation, because the MSE of our method was smaller than its counterpart, and in some cases substantially smaller. We can also see that our method converged faster.

The cross-validated errors and prediction accuracies (of the refitted models) of both methods were similar across all the settings. Overall, the simulation verified that when the considered covariates had structures corresponding to structures in the true data generating model, incorporating selection rules in variable selection can better recover the sparsity pattern and results in a better estimation. We also recommend applying the 1se rule, especially when

the goal is to recover the sparsity pattern.

5.3 Discussion

In the work presented in this chapter, we developed methods for structural learning in Cox models with time-dependent covariates, which broadens the scope of structural learning. The relevant optimization technique borrows from the work by Mairal et al. (2010), but with a different presentation, we connected the algorithm to the framework proposed in Chapter 3 and the target research question. In addition, backtracking line search was applied to the proposed method, which improves the algorithm's efficiency. The results demonstrated the benefits of employing our method over a non-structural learning approach when the structure is informative of the true model. Though no theoretical properties of overlapping group Lasso have been established yet for the Cox model, our simulation results may shed light on the future theoretical development of the proposed method.

While our method deals with selecting variables when the outcomes are time-to-event and covariates are time-varying, there are several evident avenues for future work. For instance, one could relax the assumption that the hazard is dependent on the current value of the covariates, assume event times follow a parametric distribution by using accelerated failure time models (Kalbfleisch and Prentice, 2011), generalize to different penalty types (for example, minimax concave penalty, Zhang 2010, or smoothly clipped absolute deviation, Fan and Li 2001), and/or investigate the impact of applying different weighting schemes.

The sequence of values considered for the penalization parameter λ in overlapping group Lasso needs to be set properly according to the specific question. A method to find the maximum value such that all coefficients reach 0 would be desired for automation. However, due to the complexity of the penalty, it needs more theoretical work.

More simulation can be done in the future to test 1) if the percentage of censoring plays

a role in the performance of the methods, and 2) if the method can handle the non-linear transformation of variables.

As a counterpart of overlapping group Lasso, Jacob et al. (2009) developed *latent overlapping group Lasso*, which approaches structural learning in another way by setting latent variables for a candidate variable, and then penalizing groups of latent variable coefficients. It would be interesting to extend the technique to accommodate time-dependent covariates, and compare to ours.

Chapter 6

Conclusion

6.1 Summary

This thesis focused on improving the interpretability and prediction accuracy of models resulting from variable selection. One solution is to incorporate covariate structures into variable selection. Towards this end, this thesis answered the following questions: 1) how to define the structures in variable selection, 2) how should existing methods incorporate the structures, 3) can the existing methods be used for survival data with time-dependent covariates, and 4) are there any new techniques that can incorporate a wider range of structures.

In Chapter 3, I established a theoretical framework for variable selection to formally formulate selection rules, selection dictionaries and their relationship to each other. The newly proposed concepts unify how researchers can approach variable selection problems, and encourage practitioners to incorporate their a priori knowledge in the analysis. The completeness of the framework enables practitioners to integrate any arbitrary covariate structures in variable selection. Next, I connected the framework to existing penalized regression methods, and provided theorems that can verify if a postulated grouping structure used with the

(latent) overlapping group Lasso can respect a given selection rule. The theorems inspired the construction of roadmaps for grouping structure identification in Chapters 4, which provide guidance on how to practically use the (latent) overlapping group Lasso to incorporate covariate structures. However, the existing methods cannot respect all selection rules. To overcome this issue, I showed that the task of incorporating selection rules can be formulated as optimization problems, which can be solved by MIO solvers. This enlarges the scope of covariate structures that can be incorporated into variable selection.

In Chapter 4, I showed how to use the framework and the latent overlapping group Lasso to identify predictors of major bleeding among hospitalized hypertensive patients using oral anticoagulants for atrial fibrillation. In this application, complex covariate structures arose from the relationships among the drugs of interest, the related dose and adherence to regular usage, and the drug-drug interactions, and these were integrated into the selection strategies. The selected model had a higher cross-validated risk than the model selected without the selection rules, but retained its interpretability. The results of this application will be informative for clinicians in terms of identifying which factors are more predictive of major bleeding, allowing them to identify patients at higher risk and potentially offer a more intensive or alternative follow-up. In addition, based on the theorems developed in Chapter 3, I produced roadmaps for grouping structure identification for respecting the types of selection rules used in the application, which serve as a practical guide for using latent overlapping group Lasso to incorporate covariate structures.

In Chapter 5, I focused on the overlapping group Lasso, and extended it to the time-dependent Cox model. Though the algorithm for overlapping group Lasso is fully developed, it is challenging to understand for researchers who lack the background in operational research and machine learning. To break the knowledge barrier, I presented the algorithm in an accessible form in the context of survival data. Furthermore, a simulation study was conducted to show how to practically incorporate covariate structure by using the framework

developed in 3 and the overlapping group Lasso, and the superior performance as measured by various statistics compared to the Lasso penalty (applied in the time-dependent Cox model).

6.2 Future work

To facilitate the practice of incorporating covariate structures using existing methods, I have begun developing software, including an R Shiny app, to enable the automated identification of selection dictionaries and grouping structures.

We formulated rule-based variable selection optimization problems in Chapter 3, but the theoretical properties of the associated estimators remain unknown. It is of interest to investigate those optimization problems from a theoretical and empirical perspective. This work may lead to a new paradigm of structured variable selection methods.

I have made much progress on an `R` package that can perform the structured variable selection in time-dependent Cox models from Chapter 5. Combined with the software for grouping structure identification, one would then be able to incorporate a wide range of selection rules in an application with a time-to-event outcome and time-varying covariates.

6.3 Concluding remarks

Predictive models that focus merely on prediction accuracy may lack interpretability. Therefore, users may be reluctant to use them. Increasing the interpretability of predictive model can boost confidence in their usage, and also provide mechanistic insight. We have contributed to the body of research that develops penalized regression methods with the goal of developing interpretable predictive models. Despite a large body of literature on the topic, none of the existing methods solves the problem in full generality. We therefore unified the structured variable selection problem by proposing a general framework that can integrate

universal selection rules. The framework does not only facilitate the use of existing methods, but also makes it possible to develop new classes of optimization problems with the goal of being able to respect any arbitrary selection rule.

Appendices

APPENDIX A

Appendix to Manuscript 1

A.1 Proof of Theorem 1

Proof. Suppose $\mathbb{D}_{\mathfrak{r},1}$ and $\mathbb{D}_{\mathfrak{r},2}$ respect the same selection rule $\mathfrak{r}$. Denote a subset of $\mathbb{V}$ by $\mathbb{d}$. If $\mathbb{d} \in \mathbb{D}_{\mathfrak{r},1}$, then $\mathbb{d} \in \mathbb{D}_{\mathfrak{r},2}$ by Definition 3. Without loss of generality, now suppose $\mathbb{d} \notin \mathbb{D}_{\mathfrak{r},1}$, then by Definition 3, $\mathbb{d}$ does not respect $\mathfrak{r}$, so $\mathbb{d} \notin \mathbb{D}_{\mathfrak{r},2}$. Therefore, $\mathbb{D}_{\mathfrak{r},1} = \mathbb{D}_{\mathfrak{r},2}$. Therefore, there is a unique dictionary for a given selection rule. □

A.2 Proof of Theorem 2

Proof. When $|\mathbb{F}| < \max(\mathbb{C})$ then $\mathfrak{u}_{\mathbb{C}}(\mathbb{F})$ is incoherent and the resulting unit dictionary is defined as the $\emptyset$.

When $\mathfrak{u}_{\mathbb{C}}(\mathbb{F})$ is a coherent unit rule, suppose $\mathbb{d} \in \mathcal{P}(\mathbb{V})$ respects $\mathfrak{u}_{\mathbb{C}}(\mathbb{F})$. Let $\mathbb{a} = \mathbb{d} \cap \mathbb{F} \subseteq \mathbb{F}$ such that $|\mathbb{a}| \in \mathbb{C}$. Let $\mathbb{b} = \mathbb{d} \cap (\mathbb{V} \setminus \mathbb{F})$. Then $\mathbb{d} = \mathbb{a} \cup \mathbb{b} \in \{\mathbb{a} \cup \mathbb{b}, \forall \mathbb{a} \subseteq \mathbb{F}\ s.t.\ |\mathbb{a}| \in \mathbb{C}, \forall \mathbb{b} \subseteq \mathbb{V} \setminus \mathbb{F}\}$.

Now suppose $\mathbb{d} \in \mathcal{P}(\mathbb{V})$ does not respect $\mathfrak{u}_{\mathbb{C}}(\mathbb{F})$. Then $\mathbb{d} \cap \mathbb{F}$ does not respect $\mathfrak{u}_{\mathbb{C}}(\mathbb{F})$. Necessarily, it means that $|\mathbb{d} \cap \mathbb{F}| \notin \mathbb{C}$. Therefore $\mathbb{d} \cap \mathbb{F} \notin \{\mathbb{a} \cup \mathbb{b}, \forall \mathbb{a} \subseteq \mathbb{F}\ s.t.\ |\mathbb{a}| \in \mathbb{C}, \forall \mathbb{b} \subseteq \mathbb{V} \setminus \mathbb{F}\}$, which implies $\mathbb{d} \notin \{\mathbb{a} \cup \mathbb{b}, \forall \mathbb{a} \subseteq \mathbb{F}\ s.t.\ |\mathbb{a}| \in \mathbb{C}, \forall \mathbb{b} \subseteq \mathbb{V} \setminus \mathbb{F}\}$.

Now we prove that when $\mathfrak{u}_{\mathbb{C}}(\mathbb{F})$ is a coherent unit rule, the related unit function is a bijection.

We prove by contradiction. Suppose there exists two non-empty sets $\mathbb{F}_1 \neq \mathbb{F}_2$, necessarily respecting $|\mathbb{F}_1| \geqslant \max(\mathbb{C}), |\mathbb{F}_2| \geqslant \max(\mathbb{C})$, such that $f_{\mathbb{C}}(\mathbb{F}_1) = f_{\mathbb{C}}(\mathbb{F}_2)$. Denote $\mathbb{M} = \{\mathfrak{a} \cup \mathfrak{b} : \forall \mathfrak{a} \subseteq \mathbb{F}_1 \ s.t. \ |\mathfrak{a}| \in \mathbb{C}, \forall \mathfrak{b} \subseteq \mathbb{V} \setminus \mathbb{F}_1\}$, and $\mathbb{N} = \{\mathfrak{a} \cup \mathfrak{b} : \forall \mathfrak{a} \subseteq \mathbb{F}_2 \ s.t. \ |\mathfrak{a}| \in \mathbb{C}, \forall \mathfrak{b} \subseteq \mathbb{V} \setminus \mathbb{F}_2\}$. So that $\forall \mathfrak{m} \in \mathbb{M}$, $\mathfrak{m}$ satisfies $|\mathfrak{m} \cap \mathbb{F}_1| \in \mathbb{C}$, and $\forall \mathfrak{n} \in \mathbb{N}$, $\mathfrak{n}$ satisfies $|\mathfrak{n} \cap \mathbb{F}_2| \in \mathbb{C}$. By the previous result, if $f_{\mathbb{C}}(\mathbb{F}_1) = f_{\mathbb{C}}(\mathbb{F}_2)$, then $\mathbb{M} = \mathbb{N}$. If $\mathbb{F}_1 \neq \mathbb{F}_2$, then there exists some non-empty $\mathfrak{x}$ such that $\mathfrak{x} \subseteq \mathbb{F}_1$ and $\mathfrak{x} \nsubseteq \mathbb{F}_2$. Suppose that $|\mathfrak{x}| \geqslant \min(C)$. Then $\exists \mathfrak{y}$ such that $\mathfrak{y} \subseteq \mathfrak{x}$ and $|\mathfrak{y}| = \min(\mathbb{C})$. Such $\mathfrak{y}$ is necessarily an element of $\mathbb{M}$. Because $\mathbb{M} = \mathbb{N}$, $\mathfrak{y}$ is necessarily an element of $\mathbb{N}$. According to the definition of $\mathbb{N}$, $\mathfrak{y} = \mathfrak{a}_1 \cup \mathfrak{b}_1$ where $\mathfrak{a}_1$ satisfies $\mathfrak{a}_1 \subseteq \mathbb{F}_2$ such that $|\mathfrak{a}_1| \in \mathbb{C}$, and $\mathfrak{b}_1 \subseteq \mathbb{V} \setminus \mathbb{F}_2$. So necessarily, $|\mathfrak{a}_1| = \min(\mathbb{C})$ and $\mathfrak{b}_1 = \emptyset$. This contradicts $\mathfrak{y} \nsubseteq \mathbb{F}_2$, because $\mathfrak{y} = \mathfrak{a}_1 \subseteq \mathbb{F}_2$.

Now suppose that $|\mathfrak{x}| < \min(\mathbb{C})$. Because the rule is coherent, there exists $\mathfrak{m}$ such that $\mathfrak{x} \subset \mathfrak{m} \subseteq \mathbb{F}_1$ and $|\mathfrak{m}| = \min(\mathbb{C})$. So $\mathfrak{m} \in \mathbb{M} = \mathbb{N}$. Because $\mathfrak{m} \in \mathbb{N}$, we have $\mathfrak{m} = \mathfrak{a}_2 \cup \mathfrak{b}_2$, and necessarily $|\mathfrak{a}_2| = \min(\mathbb{C})$, so $\mathfrak{b}_2 = \emptyset$ and $\mathfrak{m} = \mathfrak{a}_2 \subseteq \mathbb{F}_2$. Therefore, $\mathfrak{x} \subseteq \mathbb{F}_2$, which contradicts $\mathfrak{x} \nsubseteq \mathbb{F}_2$. □

A.3 Proof of corollary 4

Proof. Without loss of generality, suppose $\mathfrak{u}_{\mathbb{C}_1}(\mathbb{F})$ is a coherent unit rule, and $\exists c_1 \in \mathbb{C}_1$ such that $c_1 \notin \mathbb{C}_2$. By Theorem 2, $\exists \mathfrak{d} \in f_{\mathbb{C}_1}(\mathbb{F})$ such that $|\mathfrak{m} \cap \mathbb{F}| = c_1$. Then by Theorem 2, because $c_1 \notin \mathbb{C}_2$, $\mathfrak{d} \notin f_{\mathbb{C}_2}(\mathbb{F})$.

□

A.4 Proof of corollary 5

Proof. By Corollary 2, the property holds when $\mathbb{F} = \mathbb{V}$. Now suppose $\mathbb{F} \subset \mathbb{V}$. By Theorem 2, $f_{\mathbb{C}}(\mathbb{F}) = \{\mathfrak{a} \cup \mathfrak{b}, \forall \mathfrak{a} \subseteq \mathbb{F}, \forall \mathfrak{b} \subseteq \mathbb{V} \setminus \mathbb{F}\}$ when $|\mathbb{F}| \geqslant \max(\mathbb{C})$, which is $\mathcal{P}(\mathbb{V})$. Thus, $f_{\mathbb{C}}(\mathbb{F}) =$

$\mathcal{P}(\mathbb{V}), \forall \mathbb{F} \subseteq \mathbb{V}$. $\square$

A.5 Proof of mapping rules on dictionaries

Proof. For each operation on rules $\mathfrak{r}_1$ and $\mathfrak{r}_2$ with respective dictionaries $\mathbb{D}_{\mathfrak{r}_1}$ and $\mathbb{D}_{\mathfrak{r}_2}$ in Table 3.2, we prove that the rule $\mathcal{O}_{\mathfrak{r}}(\mathfrak{r}_1, \mathfrak{r}_2)$ is congruent to the operation on dictionaries in the third column.

1. $\mathcal{O}_{\mathfrak{r}}(\mathfrak{r}_1) = \neg\mathfrak{r}_1$: suppose there is a set $\mathfrak{d} \in \mathcal{P}(\mathbb{V})$ such that it does not respect $\mathfrak{r}_1$. Then by Definition 5, $\mathfrak{d} \in \mathcal{P}(\mathbb{V}) \setminus \mathbb{D}_{\mathfrak{r}_1}$. Now suppose $\mathfrak{d}$ is a set that does respect $\mathfrak{r}_1$. Then $\mathfrak{d} \in \mathbb{D}_{\mathfrak{r}_1}$, and thus $\mathfrak{d} \notin \mathcal{P}(\mathbb{V}) \setminus \mathbb{D}_{\mathfrak{r}_1}$. So, the dictionary congruent to $\neg\mathfrak{r}_1$ is $\mathcal{P}(\mathbb{V}) \setminus \mathbb{D}_{\mathfrak{r}_1}$.

2. $\mathcal{O}_{\mathfrak{r}}(\mathfrak{r}_1, \mathfrak{r}_2) = \mathfrak{r}_1 \wedge \mathfrak{r}_2$: suppose there is a set $\mathfrak{d} \in \mathcal{P}(\mathbb{V})$ such that it respects $\mathfrak{r}_1$ and $\mathfrak{r}_2$. Then by Definition 5, $\mathfrak{d} \in \mathbb{D}_{\mathfrak{r}_1} \cap \mathbb{D}_{\mathfrak{r}_2}$. Without loss of generality, now suppose $\mathfrak{d}$ is a set that does not respect $\mathfrak{r}_1$, then $\mathfrak{d} \in \mathcal{P}(\mathbb{V}) \setminus \mathbb{D}_{\mathfrak{r}_1}$, and thus $\mathfrak{d} \notin \mathbb{D}_{\mathfrak{r}_1} \cap \mathbb{D}_{\mathfrak{r}_2}$. Thus, the dictionary congruent to $\mathfrak{r}_1 \wedge \mathfrak{r}_2$ is $\mathbb{D}_{\mathfrak{r}_1} \cap \mathbb{D}_{\mathfrak{r}_2}$.

3. $\mathcal{O}_{\mathfrak{r}}(\mathfrak{r}_1, \mathfrak{r}_2) = \mathfrak{r}_1 \vee \mathfrak{r}_2$: suppose there is a set $\mathfrak{d} \in \mathcal{P}(\mathbb{V})$ such that it respects $\mathfrak{r}_1$ and/or $\mathfrak{r}_2$. Then by Definition 5, $\mathfrak{d} \in \mathbb{D}_{\mathfrak{r}_1} \cup \mathbb{D}_{\mathfrak{r}_2}$. Now suppose $\mathfrak{d}$ is a set that respects neither $\mathfrak{r}_1$ nor $\mathfrak{r}_2$, then $\mathfrak{d} \in \big(\mathcal{P}(\mathbb{V}) \setminus \mathbb{D}_{\mathfrak{r}_1}\big) \cap \big(\mathcal{P}(\mathbb{V}) \setminus \mathbb{D}_{\mathfrak{r}_2}\big)$, and thus $\mathfrak{d} \notin \mathbb{D}_{\mathfrak{r}_1} \cup \mathbb{D}_{\mathfrak{r}_2}$. Thus, the dictionary congruent to $\mathfrak{r}_1 \vee \mathfrak{r}_2$ is $\mathbb{D}_{\mathfrak{r}_1} \cup \mathbb{D}_{\mathfrak{r}_2}$.

4. $\mathcal{O}_{\mathfrak{r}}(\mathfrak{r}_1, \mathfrak{r}_2) = \mathfrak{r}_1 \to \mathfrak{r}_2$: an arbitrary set $\mathfrak{d} \in \mathcal{P}(\mathbb{V})$ falls into one of four categories, 1) $\mathfrak{d}$ respects both $\mathfrak{r}_1$ and $\mathfrak{r}_2$, 2) $\mathfrak{d}$ respects neither $\mathfrak{r}_1$ nor $\mathfrak{r}_2$, 3) $\mathfrak{d}$ respects only $\mathfrak{r}_2$ but not $\mathfrak{r}_1$, and 4) $\mathfrak{d}$ respects only $\mathfrak{r}_1$ but not $\mathfrak{r}_2$. A set $\mathfrak{d}$ in the first three categories respects $\mathfrak{r}_1 \to \mathfrak{r}_2$. We first show that sets $\mathfrak{d}$ in the first three categories belong to $(\mathcal{P}(\mathbb{V}) \setminus \mathbb{D}_{\mathfrak{r}_1}) \cup (\mathbb{D}_{\mathfrak{r}_1} \cap \mathbb{D}_{\mathfrak{r}_2})$, and a set $\mathfrak{d}$ in category 4) does not.

   (a) If $\mathfrak{d}$ is in category 1), then $\mathfrak{d} \in \mathbb{D}_{\mathfrak{r}_1} \cap \mathbb{D}_{\mathfrak{r}_2}$, which belongs to $\big(\mathcal{P}(\mathbb{V}) \setminus \mathbb{D}_{\mathfrak{r}_1}\big) \cup (\mathbb{D}_{\mathfrak{r}_1} \cap \mathbb{D}_{\mathfrak{r}_2})$.

   (b) If $\mathfrak{d}$ is in category 2), then $\mathfrak{d} \in \big(\mathcal{P}(\mathbb{V}) \setminus \mathbb{D}_{\mathfrak{r}_1}\big) \cap \big(\mathcal{P}(\mathbb{V}) \setminus \mathbb{D}_{\mathfrak{r}_2}\big)$, which belongs to $\big(\mathcal{P}(\mathbb{V}) \setminus \mathbb{D}_{\mathfrak{r}_1}\big) \cup (\mathbb{D}_{\mathfrak{r}_1} \cap \mathbb{D}_{\mathfrak{r}_2})$.

(c) If $\mathbb{d}$ is in category 3), then $\mathbb{d} \in \big(\mathcal{P}(\mathbb{V}) \setminus \mathbb{D}_{\mathfrak{r}_1}\big) \cap \mathbb{D}_{\mathfrak{r}_2}$, which belongs to $\big(\mathcal{P}(\mathbb{V}) \setminus \mathbb{D}_{\mathfrak{r}_1}\big) \cup (\mathbb{D}_{\mathfrak{r}_1} \cap \mathbb{D}_{\mathfrak{r}_2})$.

(d) If $\mathbb{d}$ is in category 4), then $\mathbb{d} \in \mathbb{D}_{\mathfrak{r}_1} \cap \big(\mathcal{P}(\mathbb{V}) \setminus \mathbb{D}_{\mathfrak{r}_2}\big)$, which does not belong to $\big(\mathcal{P}(\mathbb{V}) \setminus \mathbb{D}_{\mathfrak{r}_1}\big) \cup (\mathbb{D}_{\mathfrak{r}_1} \cap \mathbb{D}_{\mathfrak{r}_2})$.

This completes the proof.

5. This operation contains two selection steps. We have to know the result of the first selection step, $\mathbb{m} \in \mathbb{D}_{\mathfrak{r}_1}$, to obtain the resulting selection dictionary. To understand the general proof, we consider the example illustrated in the main text. Suppose $\mathbb{V} = \{A, B, C, D\}$. The first selection rule $\mathfrak{r}_1$ is "select $\{0, 2\}$ in $\{A, B\}$ and select $\{0, 2\}$ in $\{C, D\}$", and $\mathbb{D}_{\mathfrak{r}_1} = \big\{\emptyset, \{A, B\}, \{C, D\}, \{A, B, C, D\}\big\}$. The second rule $\mathfrak{r}_2$ is "if A is selected, then B must be selected, and if C is selected, then D must be selected", so $\mathbb{D}_{\mathfrak{r}_2} = \big\{\emptyset, \{B\}, \{A, B\}, \{D\}, \{C, D\}, \{B, D\}, \{B, C, D\}, \{A, B, D\}, \{A, B, C, D\}\big\}$.

The operation is $\mathfrak{r}_1 \Rightarrow \mathfrak{r}_2$. Suppose that the first step selects $\mathbb{m} = \{A, B\}$. The resulting dictionary that is congruent to $\mathfrak{r}_1 \Rightarrow \mathfrak{r}_2$ is denoted $\mathbb{A}_{\mathbb{m}} = \{\mathbb{a} : \mathbb{a} \in \mathbb{D}_{\mathfrak{r}_2}, \mathbb{a} \subseteq \mathbb{m}\}$, which is a set of sets, where each set must be 1) an element in $\mathbb{D}_{\mathfrak{r}_2}$ and 2) be a subset of $\mathbb{m}$. In the example we get $\mathbb{A}_{\mathbb{m}} = \{\emptyset, \{B\}, \{A, B\}\}$.

Now we prove it formally. As before, let $\mathbb{m}$ be the result of a selection step respecting $\mathfrak{r}_1$ and $\mathbb{A}_{\mathbb{m}}$ is defined as above. Suppose $\mathbb{d}$ respects $\mathfrak{r}_1 \Rightarrow \mathfrak{r}_2$. Then $\mathbb{d}$ must respect $\mathbb{D}_{\mathfrak{r}_2}$, and $\mathbb{d}$ must also be a subset of $\mathbb{m}$ because the second selection step follows the first one that selected $\mathbb{m}$. Now suppose $\mathbb{d}$ does not respect $\mathfrak{r}_1 \Rightarrow \mathfrak{r}_2$, then $\mathbb{d}$ must satisfy one of the following conditions: 1) $\mathfrak{r}_1$ is not being respected, regardless of whether $\mathfrak{r}_2$ is being respected, so $\mathbb{d} \in \mathcal{P}(\mathbb{V}) \setminus \mathbb{D}_{\mathfrak{r}_1}$. 2) $\mathfrak{r}_1$ is being respected, but given the resulting set of selected variables $\mathbb{m}$, $\mathfrak{r}_2$ is not being respected, then $\mathbb{d} \in \{\mathbb{b} : \mathbb{b} \subseteq \mathbb{m}, \mathbb{b} \notin \mathbb{D}_{\mathfrak{r}_2}\}$. Any $\mathbb{d}$ satisfying the above conditions does not belong to $\mathbb{A}_{\mathbb{m}}$, which completes the proof.

□

A.6 Proof of Theorem 3

Proof. The theorem is equivalent to saying that for a given rule $\mathfrak{r}$ on $\mathbb{V}$, the related dictionary $\mathbb{D}$ can be obtained by unions and/or intersections of unit dictionaries.

Suppose that the selection dictionary has cardinality 0. Then it is equal to a unit dictionary of an incoherent unit rule.

Now suppose that the selection dictionary is a set with cardinality 1. Let $\mathbb{D}_{\mathfrak{r}} = \{\mathbb{F}\}$, for some $\mathbb{F} \subseteq \mathbb{V}$. Let $\mathbb{D}_{\mathfrak{u}_1}$ and $\mathbb{D}_{\mathfrak{u}_2}$ be dictionaries corresponding to unit rules $\mathfrak{u}_1 = \mathfrak{u}_{\{|\mathbb{F}|\}}(\mathbb{F})$ and $\mathfrak{u}_2 = \mathfrak{u}_{\{0\}}(\mathbb{V} \setminus \mathbb{F})$, respectively. Then $\mathbb{D}_{\mathfrak{r}}$ can be expressed as $\mathbb{D}_{\mathfrak{u}_1} \cap \mathbb{D}_{\mathfrak{u}_2}$. Thus, $\mathfrak{r} = \mathfrak{u}_1 \wedge \mathfrak{u}_2$.

We have demonstrated that we can construct a selection dictionary with a single element using unit dictionaries. Selection dictionaries containing more than one element can be constructed by taking the unions of selection dictionaries with single elements. □

A.7 Proof of Theorem 4

Proof. The variables being selected by latent overlapping group Lasso (LOGL) is the union of groups of variables whose latent coefficients are estimated as non-zero. (Obozinski et al., 2011a) Mathematically, let $\text{supp}(\hat{\mathbb{V}})$ be the variables being selected by LOGL, and let $\mathbb{Q} \subseteq \mathbb{G}$ be the set of groups $\mathbb{g}$ with the estimate of latent coefficients $\hat{\boldsymbol{\alpha}}_{|\mathbb{g}}$ such that $\hat{\boldsymbol{\alpha}}_{|\mathbb{g}} \neq \mathbf{0}$. Then $\text{supp}(\hat{\mathbb{V}}) = \cup_{\mathbb{g} \in \mathbb{Q}} \mathbb{g}$. Note that $\mathbb{Q}$ is an element of the power set of $\mathbb{G}$.

For a given $\mathbb{V}$ with grouping structure $\mathbb{G} := \{\mathbb{g}_i, i = 1, \ldots, I\}$, by definition of grouping structure $\cup_{i=1}^{I} \mathbb{g}_i = \mathbb{V}$. Depending on the estimation, there are 2^I possible combinations of groups with non-zero latent coefficients $\hat{\boldsymbol{\alpha}}_{|\mathbb{g}}$. That is, there are 2^I possible $\mathbb{Q}$s. List these as $\mathbb{Q}_j, j = 1, ..., 2^I$. Given $\mathbb{Q}_j$, the variables that are selected into the model by LOGL are the covariates in [any of the groups in] $\mathbb{Q}_j$. We can represent these covariates by $\mathbb{d}_j = \cup_{\mathbb{g} \in \mathbb{Q}_j} \mathbb{g}$. By definition, each $\mathbb{d}_j, j = 1, ..., 2^I$ is an element of the dictionary, i.e. one possible set of covariates that can be selected. So the congruent dictionary under this penalization structure

$(\mathcal{M}, \mathbb{G})$ is $\mathbb{D}_{\mathbb{G}}^{\mathcal{M}} = \{\mathbb{d}_j, j = 1, ..., 2^I\}$. So any selection rule with congruent dictionary equal to $\mathbb{D}_{\mathbb{G}}^{\mathcal{M}}$ is congruent to the penalization structure $(\mathcal{M}, \mathbb{G})$. Every step is necessary and sufficient so this completes the proof.

□

A.8 Theorem for overlapping group Lasso

THEOREM 5. *For a given $\mathbb{V}$, with overlapping group Lasso, the necessary condition of a grouping structure $\mathbb{G} \coloneqq \{\mathbb{g}_i, i = 1, \ldots, I\}$ being congruent to a selection rule $\mathfrak{r}$ is $\mathbb{D}_{\mathfrak{r}} \setminus \mathbb{V} = \{(\cup_{\mathbb{g} \in \mathbb{Q}_j} \mathbb{g})^{\mathsf{c}}, j = 1, \ldots, 2^I\}$, where $\mathbb{Q}_j, j = 1, ..., 2^I$ are all unique subsets of $\mathbb{G}$.*

APPENDIX B

Appendix to Manuscript 2

B.1 Population-based cohort definition flowchart

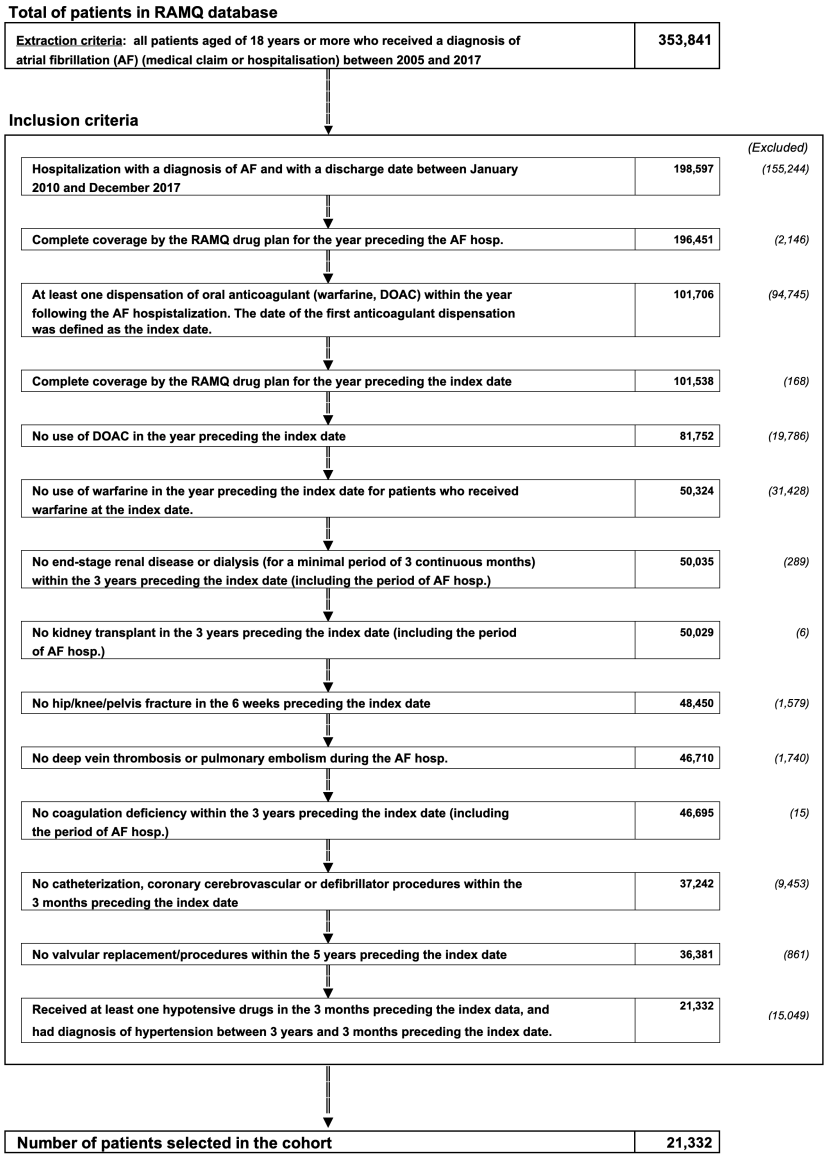


Figure B.1: Population-based cohort definition flowchart.

Penalty	Specification
L^2	$f(\boldsymbol{x};\boldsymbol{\lambda}) = \lambda \left\|\boldsymbol{x}\right\|_2$
MCP	$f(\boldsymbol{x};\boldsymbol{\lambda}) = \begin{cases} \lambda\lvert\boldsymbol{x}\rvert - \frac{\boldsymbol{x}^2}{2\gamma}, & \text{if } \lvert\boldsymbol{x}\rvert \leqslant \gamma\lambda \\ \frac{1}{2}\gamma\lambda^2, & \text{if } \lvert\boldsymbol{x}\rvert > \gamma\lambda \end{cases} \quad \gamma > 1$
SCAD	$f(\boldsymbol{x};\boldsymbol{\lambda}) = \begin{cases} \lambda\lvert\boldsymbol{x}\rvert, & \text{if } \lvert\boldsymbol{x}\rvert \leqslant \lambda, \\ \frac{2\gamma\lambda\lvert\boldsymbol{x}\rvert - \boldsymbol{x}^2 - \lambda^2}{2(\gamma-1)}, & \text{if } \lambda < \lvert\boldsymbol{x}\rvert < \gamma\lambda, \quad \gamma > 2 \\ \frac{\lambda^2(\gamma+1)}{2}, & \text{if } \lvert\boldsymbol{x}\rvert \geqslant \gamma\lambda \end{cases}$

Table B.1: Different specification of penalties. In MCP and SCAD, $\boldsymbol{x}$ should be replaced by $\left\|\boldsymbol{\alpha}^{\mathbb{g}_i}\right\|_2$.

B.2 Different specification of penalties

B.3 More details of the roadmaps

There are more than one grouping structure that can respect a same selection rule used with the latent overlapping group lasso. The roadmap provided in Table 4.2 shows the way to identify the most efficient grouping structure in the sense that the grouping structure contains the least number of groups among all eligible grouping structures.

Within the interaction selection application, when all the variables in $\mathbb{A}$ are the interactions of the variables in $\mathbb{B}$, selection rule 1 and 2 in Table 4.2 correspond to strong and weak heredity respectively. (Haris et al., 2016a) When $n = 1$ and $m = 2$, the two selection rules degrade to two-way interaction selection. In this simple case, the grouping structure identification was mentioned by (Yan et al., 2017).

Note that the roadmap works with conditions: all variables in $\mathbb{A}$ and $\mathbb{B}$ are continuous or binary variables, and $\mathbb{A} \cap \mathbb{B} = \emptyset$. When there are categorical variables in $\mathbb{B}$, and thus $\mathbb{A}$, one should be more careful to specify the grouping structure in the sense that it is necessary to group the dummy variables in $\mathbb{B}$ representing a categorical variable, but not the ones in $\mathbb{A}$. In addition, the grouping structure desires to be specified case by case when there are more than one rules being applied in a set of variables. For example, if the selection rule is "if

A is selected, the {B, C} must be selected" and "if D is selected, then $\{A, B, C\}$ must be selected". Then the grouping structure should be $\big\{\{B\}, \{C\}, \{A, B, C\}, \{A, B, C, D\}\big\}$. $\{A\}$ should not be a single group because it cannot be selected alone.

B.4 Rationale of the modification of rule 3

The the latent overlapping group lasso cannot practically respect the selection rule 3: all the 10 variables must be selected. This is because, first, the latent overlapping group lasso requires that all variables must belong to at least one group in the penalty term: all coefficients of variables must be penalized at a certain level. That is, once a group is specified, the group of variables are possibly not being selected theoretically. Second, since the weight $\omega_{\mathfrak{g}}$ must be positive, one may attempt to set the weight of the group to a fairly small number close to 0, which is effectively the same as no penalization on those 10 variables. However, the estimates of the latent overlapping group lasso are very sensitive to the weights specification. A zero weight for a group would ruin the estimates, see details in (Obozinski et al., 2011a). We set the rule 3 because from the literature we know that these variables must be predictive. With this knowledge, the modified rule is effectively the same as the original selection rule, because it would be impossible to select none of those variables from the application.

B.5 Grouping structure in the application

Define

$\mathfrak{g}_1$={CHA_2DS_2-VASc Score},

$\mathfrak{g}_2$={Heart Disease},

$\mathfrak{g}_3$={Diabetes},

$\mathfrak{g}_4$={COPD},

$\mathfrak{g}_5$={Dyslipidemia},

$\mathfrak{g}_6$={OAC type},

$\mathfrak{g}_7$={OAC type, Dose},

$\mathfrak{g}_8$={OAC type, Apixaban},

$\mathfrak{g}_9$={OAC type, Dabigatran},

$\mathfrak{g}_{10}$={Adherence},

$\mathfrak{g}_{11}$={OAC type, Adherence, Interaction of OAC type and Adherence},

$\mathfrak{g}_{12}$={OAC type, Adherence, Interaction of OAC type and Adherence, Dose, Interaction of Dose and Adherence},

$\mathfrak{g}_{13}$={OAC type, Interaction of OAC type and Antiplatelets},

$\mathfrak{g}_{14}$={OAC type, Interaction of OAC type and NSAIDs},

$\mathfrak{g}_{15}$={Antidepressants},

$\mathfrak{g}_{16}$={OAC type, Antidepressants, Interaction of OAC type and Antidepressants},

$\mathfrak{g}_{17}$={PPI},

$\mathfrak{g}_{18}$={OAC type, PPI, Interaction of OAC type and PPI},

$\mathfrak{g}_{19}$={Age, History of major bleeding, Stroke, Anemia, Sex, Renal diseases, Liver disease, Malignancy, Antiplatelets, NSAIDs}.

The grouping structure is the $\mathbb{G} = \{\mathfrak{g}_i, i = 1, \ldots, 19\}$

B.6 Demographic and characteristics of patients stratified by DOAC dose and warfarin

	DOAC		
Variable name (non-reference/reference level)	**High dose**	**Low dose**	**warfarin**
	Mean (SD)	Mean (SD)	Mean (SD)
	n=7022 (32%)	n=5067 (24%)	n=9243 (43%)
Baseline covariates			
1. Age ($\geqslant$75/<75)	0.55 (0.50)	0.91 (0.29)	0.77 (0.42)
2. Sex (female/male)	0.50 (0.50)	0.66 (0.47)	0.59 (0.49)

3. CHA_2DS_2-VASc Score ($\geqslant 3/<3$)	0.79 (0.41)	0.95 (0.22)	0.91 (0.29)
Comorbidities within 3 years before cohort entry			
4. Stroke (yes/no)	0.25 (0.66)	0.27 (0.69)	0.30 (0.72)
5. Anemia (yes/no)	0.06 (0.24)	0.09 (0.29)	0.13 (0.34)
6. Malignancy (yes/no)	0.27 (0.44)	0.26 (0.44)	0.26 (0.44)
7. Liver disease (yes/no)	0.02 (0.15)	0.02 (0.13)	0.02 (0.15)
8. History of major bleeding (yes/no)	0.28 (0.45)	0.36 (0.48)	0.37 (0.48)
9. Renal diseases (yes/no)	0.18 (0.38)	0.27 (0.45)	0.35 (0.48)
10. Heart disease (yes/no)	0.55 (0.50)	0.65 (0.48)	0.69 (0.46)
11. Diabetes (yes/no)	0.36 (0.48)	0.30 (0.46)	0.40 (0.49)
12. COPD/asthma (yes/no)	0.38 (0.49)	0.37 (0.48)	0.41 (0.49)
13. Dyslipidemia (yes/no)	0.59 (0.49)	0.55 (0.50)	0.59 (0.49)
OAC use at cohort entry			
14. OAC type (DOAC/warfarin)	1.00 (0.00)	1.00 (0.00)	0.00 (0.00)
15. Apixaban	0.54 (0.50)	0.50 (0.50)	0.00 (0.00)
16. Dabigatran	0.11 (0.31)	0.28 (0.45)	0.00 (0.00)
17. Dose (high dose DOAC/low dose DOAC or warfarin)	1.00 (0.00)	0.00 (0.00)	0.00 (0.00)
18. Adherence (high/low)	0.85 (0.35)	0.86 (0.35)	0.88 (0.32)
19. Interaction of OAC type and Adherence	0.85 (0.35)	0.86 (0.35)	0.00 (0.00)
20. Interaction of Dose and Adherence	0.85 (0.35)	0.00 (0.00)	0.00 (0.00)
Concomitant medication use within 2 weeks before cohort entry			
21. Antiplatelets (yes/no)	0.27 (0.44)	0.32 (0.47)	0.37 (0.48)
22. NSAIDs (yes/no)	0.01 (0.12)	0.01 (0.10)	0.01 (0.10)
23. Antidepressants (yes/no)	0.18 (0.38)	0.20 (0.40)	0.18 (0.38)
24. PPIs (Proton pump inhibitors) (yes/no)	0.37 (0.48)	0.44 (0.50)	0.49 (0.50)
Potential drug-drug interaction			
25. Interaction of OAC type and Antiplatelets	0.27 (0.44)	0.32 (0.47)	0.00 (0.00)
26. Interaction of OAC type and NSAIDs	0.01 (0.12)	0.01 (0.10)	0.00 (0.00)
27. Interaction of OAC type and Antidepressants	0.18 (0.38)	0.20 (0.40)	0.00 (0.00)
28. Interaction of OAC type and PPIs	0.37 (0.48)	0.44 (0.50)	0.00 (0.00)
Outcome			

29. Major bleeding	0.02 (0.15)	0.03 (0.18)	0.04 (0.19)

Table B.2: Covariates descriptive statistics stratified by Dose

B.7 Crude and adjusted odds ratios and 95% confidence intervals of variables

Variable name (non-reference/reference level)	**Crude OR (95% CI)**	**Adjusted OR (95% CI)**
Elderly (>65/⩽ 65)	2.51 (1.53, 4.48)	2.43 (1.48, 4.35)
History of major bleeding (yes/no)	1.87 (1.60, 2.19)	1.78 (1.51, 2.08)
Liver diseases (yes/no)	2.35 (1.60, 3.34)	2.15 (1.46, 3.06)
Renal diseases (yes/no)	1.36 (1.15, 1.60)	1.14 (0.96, 1.34)
Stroke (yes/no)	0.96 (0.85, 1.07)	0.93 (0.82, 1.04)
Drugs* (yes/no)	1.23 (1.04, 1.45)	1.18 (1.00, 1.40)

Table B.3: Crude (univariate) and adjusted odds ratios from simple and multivariate logistic regression models, respectively and 95% confidence intervals using the variables in HAS-BLED chart (2012, Circulation, Lane at el.)

* Drugs: a binary variable, it is 1 if the patient had at least 1 dispensation of the following drugs in the 3 years preceding the index date or during the AF hospitalization: Clopidogrel, low dose of ASA (daily dose < 100 mg), NSAID.

Variable name (non-reference/reference level)	**Crude OR (95% CI)**	**Adjusted OR (95% CI)**
Age ($\geqslant$75/<75)	1.28 (1.07, 1.55)	1.29 (1.07, 1.56)
Sex (female/male)	0.84 (0.72, 0.98)	0.86 (0.73, 1.01)
Adherence (high/low)	1.04 (0.83, 1.32)	1.03 (0.82, 1.32)
History of major bleeding (yes/no)	1.87 (1.60, 2.19)	1.68 (1.43, 1.98)
Renal diseases (yes/no)	1.36 (1.15, 1.60)	1.05 (0.88, 1.25)
Liver diseases (yes/no)	2.35 (1.60, 3.34)	2.09 (1.42, 2.99)
Stroke (yes/no)	0.96 (0.85, 1.07)	0.94 (0.83, 1.05)
Anemia (yes/no)	1.86 (1.50, 2.28)	1.46 (1.16, 1.81)
Malignancy (yes/no)	1.21 (1.02, 1.43)	1.09 (0.91, 1.29)
Antiplatelets (yes/no)	1.40 (1.19, 1.64)	1.31 (1.11, 1.54)
NSAIDs (yes/no)	1.15 (0.54, 2.11)	1.29 (0.61, 2.38)
Other medications* (yes/no)	1.22 (1.03, 1.46)	1.06 (0.88, 1.27)

Table B.4: Crude (univariate) and adjusted odds ratios from simple and multivariate logistic regression models, respectively and 95% confidence intervals using the risk factors of bleeding on Oral anticoagulation (2012, Circulation, Lane at el.)

Other medications: a binary variable, it is 1 if the patient had at least 1 dispensation of the following drugs within the 2 weeks prior to index date: Antidiabetics, Antidepressants, PPIs, Antibiotics, Antiarrhythmics, PGP inhibitors.

Variable name (non-reference/reference level)	Crude OR (95% CI)	Adjusted OR (95% CI)
Baseline covariates		
1. Age (⩾75/<75)	1.28 (1.07, 1.55)	1.23 (1.00, 1.54)
2. Sex (female/male)	0.84 (0.72, 0.98)	0.86 (0.73, 1.02)
3. CHA_2DS_2VASc score (⩾3/<3)	1.54 (1.17, 2.06)	1.24 (0.90, 1.75)
Comorbidities within 3 years before cohort entry		
4. Stroke (yes/no)	0.96 (0.85, 1.07)	0.94 (0.83, 1.05)
5. Anemia (yes/no)	1.86 (1.50, 2.28)	1.38 (1.10, 1.72)
6. Malignancy (yes/no)	1.21 (1.02, 1.43)	1.09 (0.91, 1.30)
7. Liver diseases (yes/no) (yes/no)	2.35 (1.60, 3.34)	2.01 (1.36, 2.89)
8. History of major bleeding (yes/no)	1.87 (1.60, 2.19)	1.62 (1.37, 1.91)
9. Renal diseases (yes/no)	1.36 (1.15, 1.60)	0.96 (0.80, 1.15)
10. Heart disease (yes/no)	1.51 (1.27, 1.80)	1.23 (1.02, 1.48)
11. Diabetes (yes/no)	1.49 (1.27, 1.74)	1.30 (1.10, 1.54)
12. COPD/asthma (yes/no)	1.33 (1.14, 1.55)	1.15 (0.98, 1.35)
13. Dyslipidemia (yes/no)	1.28 (1.09, 1.50)	1.09 (0.92, 1.30)
OAC use at cohort entry		
14. OAC type (DOACs/warfarin)	0.76 (0.65, 0.89)	1.69 (0.94, 2.98)
15. Apixaban (yes/no), ref: Rivaroxaban	0.63 (0.52, 0.76)	0.64 (0.50, 0.82)
16. Dabigatran (yes/no), ref: Rivaroxaban	1.17 (0.91, 1.47)	1.07 (0.78, 1.45)
17. Dose (high dose DOACs/low dose DOACs or warfarin)	0.68 (0.56, 0.81)	0.66 (0.36, 1.19)
18. Adherence (high/low)	1.04 (0.83, 1.32)	1.02 (0.72, 1.48)
19. Interaction of OAC type and Adherence	0.79 (0.67, 0.92)	0.82 (0.47, 1.45)
20. Interaction of Dose and Adherence	0.71 (0.58, 0.85)	1.29 (0.69, 2.42)
Concomitant medication use within 2 weeks before cohort entry		
21. Antiplatelets (yes/no)	1.40 (1.19, 1.64)	1.48 (1.18, 1.86)
22. NSAIDs (yes/no)	1.15 (0.54, 2.11)	1.03 (0.25, 2.78)
23. Antidepressants (yes/no)	1.27 (1.05, 1.52)	1.22 (0.93, 1.59)
24. PPIs (yes/no)	1.03 (0.88, 1.20)	0.85 (0.68, 1.07)
Potential drug-drug interaction		
25. Interaction of OAC type and Antiplatelets	0.95 (0.76, 1.17)	0.71 (0.51, 0.99)

26. Interaction of OAC type and NSAIDs	1.26 (0.49, 2.62)	1.52 (0.38, 7.42)
27. Interaction of OAC type and Antidepressants	0.98 (0.75, 1.26)	0.90 (0.61, 1.32)
28. Interaction of OAC type and PPIs	0.78 (0.64, 0.95)	0.88 (0.63, 1.22)

Table B.5: Crude (univariate) and adjusted odds ratios from simple and multivariate logistic regression models, respectively and 95% confidence intervals using the covariates in in the analysis

B.8 Derivation of estimated odds ratios of taking different types of DOACs versus warfarin from the selected models

We focus on the selected model resulting from the latent overlapping group lasso using MCP/SCAD penalty. Based on the current variable definitions, we show the corresponding variable values when the patient took different type of OACs below.

Category/Variable name	OAC Type	Dose	Apixaban	Dabigatran
High dose Apixaban	1	1	1	0
High dose Dabigatran	1	1	0	1
High dose Rivaroxaban	1	1	0	0
Low dose Apixaban	1	0	1	0
Low dose Dabigatran	1	0	0	1
Low dose Rivaroxaban	1	0	0	0
warfarin	0	0	0	0

Table B.6: Variable values when the patient took different types of OACs (category)

Suppose we fit a logistic regression

$$logit\{E(Y)\} = \beta_0 + \beta_1 \text{OAC Type} + \beta_2 \text{Dose} + \beta_3 \text{Apixaban} + \beta_4 \text{Dabigatran} + ...,$$

where Y is the outcome, major bleeding. For brevity, we omit the other variables. Then the estimated probabilities of major bleeding for different types of OACs are given in Table B.7.

Subject who took...	Estimates
High dose Apixaban	$\text{ilogit}(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 + \dots)$
High dose Dabigatran	$\text{ilogit}(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_4 + \dots)$
High dose Rivaroxaban	$\text{ilogit}(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \dots)$
Low dose Apixaban	$\text{ilogit}(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_3 + \dots)$
Low dose Dabigatran	$\text{ilogit}(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_4 + \dots)$
Low dose Rivaroxaban	$\text{ilogit}(\hat{\beta}_0 + \hat{\beta}_1 + \dots)$
warfarin	$\text{ilogit}(\hat{\beta}_0 + \dots)$

Table B.7: Estimated mean outcomes
*ilogit: inverse logit

Thus, the interpretation of the parameters are given in Table B.8.

Contrasts	Parameter(s)	Estimated odds ratios
High dose Apixaban	$\exp(\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3)$	0.70
High dose Dabigatran	$\exp(\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_4)$	1.13
High dose Rivaroxaban	$\exp(\hat{\beta}_1 + \hat{\beta}_2)$	1.13
Low dose Apixaban	$\exp(\hat{\beta}_1 + \hat{\beta}_3)$	0.86
Low dose Dabigatran	$\exp(\hat{\beta}_1 + \hat{\beta}_4)$	1.39
Low dose Rivaroxaban	$\exp(\hat{\beta}_1)$	1.39

Table B.8: Estimated odds ratios of taking different types of DOACs versus warfarin from the model selected by the latent overlapping group lasso MCP/SCAD

APPENDIX C

Appendix to Chapter 5

C.1 Proof of Lemma 1

Let us consider the conic duality (Boyd and Vandenberghe, 2004). First we introduce the cone defined on two variables $\boldsymbol{\beta}$ and z: $\mathcal{C} \coloneqq \{(\boldsymbol{\beta}, z) \in \mathbb{R}^{p+1}; \|\boldsymbol{\beta}\| \leqslant z\}$. Next we can rewrite the problem (5.1) using an additional primary variable $\boldsymbol{z} = (z_{\mathbb{g}})_{\mathbb{g}\in\mathbb{G}} \in \mathbb{R}^{|\mathbb{G}|}$ that satisfies $|\mathbb{G}|$ conic constraints $(\boldsymbol{\beta}_{|\mathbb{g}}, z_{\mathbb{g}}) \in \mathcal{C}$, for $\mathbb{g} \in \mathbb{G}$.

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p, \boldsymbol{z}\in\mathbb{R}^{|\mathbb{G}|}} \frac{1}{2t} \left\| \boldsymbol{\beta} - \{\tilde{\boldsymbol{\beta}} - t\nabla f(\tilde{\boldsymbol{\beta}})\} \right\| + \lambda \sum_{\mathbb{g}\in\mathbb{G}} \omega_{\mathbb{g}} z_{\mathbb{g}}, \text{ s.t.} (\boldsymbol{\beta}_{|\mathbb{g}}, z_{\mathbb{g}}) \in \mathcal{C}, \forall \mathbb{g} \in \mathbb{G}.$$

We can then convert the above optimization problem into a dual problem by introducing the dual variables $\boldsymbol{\tau} = (\tau_{\mathbb{g}})_{\mathbb{g}\in\mathbb{G}} \in \mathbb{R}^{|\mathbb{G}|}$, $\boldsymbol{\xi} = (\boldsymbol{\xi}_{|\mathbb{g}})_{\mathbb{g}\in\mathbb{G}} \in \mathbb{R}^{p\times|\mathbb{G}|}$, and $\mathcal{C}$'s dual counterpart $\mathcal{C}^* \coloneqq \{(\vec{\boldsymbol{\xi}}, \tau) \in \mathbb{R}^{p+1}; \|\vec{\boldsymbol{\xi}}\|_* \leqslant \tau\}$. The generalized conic inequalities also can be applied on both $\mathcal{C}$ and $\mathcal{C}^*$, thus the strong duality holds because the primary problem is convex and satisfies Slater's conditions (Boyd and Vandenberghe, 2004; Jenatton et al., 2011b). Consider

the Lagrangian

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{\tau}, \boldsymbol{\xi}) = \frac{1}{2t} \left\| \boldsymbol{\beta} - \{\tilde{\boldsymbol{\beta}} - t\nabla f(\tilde{\boldsymbol{\beta}})\} \right\|_2^2 + \lambda \sum_{\mathbb{g} \in \mathbb{G}} \omega_{\mathbb{g}} z_{\mathbb{g}} - \sum_{\mathbb{g} \in \mathbb{G}} \begin{pmatrix} z_{\mathbb{g}} \\ \boldsymbol{\beta}_{|\mathbb{g}} \end{pmatrix}^\top \begin{pmatrix} \tau_{\mathbb{g}} \\ \boldsymbol{\xi}_{|\mathbb{g}} \end{pmatrix}. \quad \text{(C.1)}$$

To obtain the dual function, we minimize out the primary variables by taking derivatives of $\mathcal{L}$ with respect to the primary variables $\boldsymbol{\beta}$ and $\boldsymbol{z}$ respectively and setting the derivatives to zeros, which gives us

$$\boldsymbol{\beta} - \{\tilde{\boldsymbol{\beta}} - t\nabla f(\tilde{\boldsymbol{\beta}})\} - \sum_{\mathbb{g} \in \mathbb{G}} \boldsymbol{\xi}_{|\mathbb{g}} = 0$$

$$\lambda \omega_{\mathbb{g}} = \tau_{\mathbb{g}}, \forall \mathbb{g} \in \mathbb{G},$$

and thus the dual function (5.3) can be obtained by plugging in the dual variable into (C.1) and flipping (without loss of generality) the sign of $\boldsymbol{\xi}$.

Next, we derive the optimality conditions from the Karush-Kuhn-Tucher conditions. We have that $\{\boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{\tau}, \boldsymbol{\xi}\}$ achieve the optimal values if (again, without loss of generality, we flip the sign of $\boldsymbol{\xi}$)

$$\forall \mathbb{g} \in \mathbb{G}, z_{\mathbb{g}} \tau_{\mathbb{g}} - \boldsymbol{\beta}_{|\mathbb{g}}^\intercal \boldsymbol{\xi}_{|\mathbb{g}} = 0, \quad \text{(C.2)}$$

$$\forall \mathbb{g} \in \mathbb{G}, (\boldsymbol{\beta}_{|\mathbb{g}}, z_{\mathbb{g}}) \in \mathcal{C}, \quad \forall \mathbb{g} \in \mathbb{G}, \lambda \omega_{\mathbb{g}} - \tau_{\mathbb{g}} = 0, \quad \text{(C.3)}$$

$$\forall \mathbb{g} \in \mathbb{G}, (\boldsymbol{\xi}_{|\mathbb{g}}, \tau_{\mathbb{g}}) \in \mathcal{C}^*, \quad \forall \mathbb{g} \in \mathbb{G}, \boldsymbol{\beta} - (\tilde{\boldsymbol{\beta}} - t\nabla f(\tilde{\boldsymbol{\beta}}) + \sum_{\mathbb{g} \in \mathbb{G}} \boldsymbol{\xi}_{|\mathbb{g}} = 0. \quad \text{(C.4)}$$

Therefore, we have

$$\begin{aligned}
\forall \mathbb{g} \in \mathbb{G}, \lambda z_{\mathbb{g}} \omega_{\mathbb{g}} =& \boldsymbol{\beta}_{|\mathbb{g}}^{\intercal} \boldsymbol{\xi}_{|\mathbb{g}} && \text{by (C.2)\&(C.3)} \\
\leqslant& \left\| \boldsymbol{\beta}_{|\mathbb{g}} \right\| \left\| \boldsymbol{\xi}_{|\mathbb{g}} \right\|_* && \text{by the definition of dual-norm} \\
\leqslant& z_{\mathbb{g}} \left\| \boldsymbol{\xi}_{|\mathbb{g}} \right\|_* && \text{by conic inequality} \\
\leqslant& \lambda z_{\mathbb{g}} \omega_{\mathbb{g}}. && \text{by the definition of dual-norm\&(C.3)}
\end{aligned}$$

From the above equation/inequalities, we have that $\boldsymbol{\beta}_{|\mathbb{g}}^{\intercal} \boldsymbol{\xi}_{|\mathbb{g}} = \left\| \boldsymbol{\beta}_{|\mathbb{g}} \right\| \left\| \boldsymbol{\xi}_{|\mathbb{g}} \right\|_*$, and $z_{\mathbb{g}} \left\| \boldsymbol{\xi}_{|\mathbb{g}} \right\|_* = \lambda z_{\mathbb{g}} \omega_{\mathbb{g}}$. We know that due to the conic inequality, if $\boldsymbol{\beta}_{|\mathbb{g}} \neq 0$, then $z_{\mathbb{g}} \neq 0$, which implies $\left\| \boldsymbol{\xi}_{|\mathbb{g}} \right\|_* = \lambda \omega_{\mathbb{g}}$. Combine those two equations, we have

$$\boldsymbol{\xi}_{|\mathbb{g}} = \begin{cases} \lambda \omega_{\mathbb{g}} \left\| \boldsymbol{\beta}_{|\mathbb{g}} \right\| (\boldsymbol{\beta}_{|\mathbb{g}})^{-1}, & \text{if } \boldsymbol{\beta}_{|\mathbb{g}} \neq 0 \\ \boldsymbol{\xi}_{|\mathbb{g}}, & \text{if } \boldsymbol{\beta}_{|\mathbb{g}} = 0. \end{cases}$$

We can also rewrite the above equation using the projection concept, which is called the projection on a dual ball

$$\boldsymbol{\xi}_{|\mathbb{g}} = \prod_{\|\cdot\|_* \leqslant \lambda \omega_{\mathbb{g}}} (\boldsymbol{\beta}_{|\mathbb{g}} + \boldsymbol{\xi}_{|\mathbb{g}}) = \prod_{\|\cdot\|_* \leqslant \lambda \omega_{\mathbb{g}}} \left\{ \left[\tilde{\boldsymbol{\beta}} - t \nabla f(\tilde{\boldsymbol{\beta}}) - \sum_{\mathbb{h} \neq \mathbb{g}} \boldsymbol{\xi}_{|\mathbb{h}} \right]_{|\mathbb{g}} \right\}, \tag{C.5}$$

where the definition of projection on a dual ball can be found in (Jenatton et al., 2011b; Borwein and Lewis, 2010): Let $\boldsymbol{w} \in \mathbb{R}^p$ and $t > 0$. We express $\boldsymbol{\kappa}$ as the projection of $\boldsymbol{w}$ on the dull norm with the length t, and

$$\boldsymbol{\kappa} = \prod_{\|\cdot\|_* \leqslant t} (\boldsymbol{w}) = \begin{cases} \boldsymbol{w} & \text{if } \|\boldsymbol{w}\|_* \leqslant t, \\ \boldsymbol{\kappa} : \|\boldsymbol{\kappa}\|_* = t, \text{and } \boldsymbol{\kappa}^{\top}(\boldsymbol{w} - \boldsymbol{\kappa}) = \|\boldsymbol{\kappa}\|_* \|\boldsymbol{w} - \boldsymbol{\kappa}\| & \text{otherwise.} \end{cases} \tag{C.6}$$

We can see from the definition of projection onto a dual ball (C.6) that, if the dual norm of the vector/function to be projected is less than the radius of the dual ball, then the projection

is itself; otherwise, it is the maximum vector/function that satisfies two conditions: 1) the dual norm is the length of the radius, and 2) for any $\|\boldsymbol{w}-\boldsymbol{\kappa}\| \leqslant 1$, the dual norm of $\boldsymbol{\kappa}$ achieves the value that maximizes $\boldsymbol{\kappa}^{\top}(\boldsymbol{w}-\boldsymbol{\kappa})$.

In this way, we can then see $\boldsymbol{\xi}_{|\mathbb{g}}$ as the projection of $\boldsymbol{\beta}_{|\mathbb{g}}+\boldsymbol{\xi}_{|\mathbb{g}}$ onto the dual ball with the radius $\lambda\omega_{\mathbb{g}}$ of the dual norm $\|\cdot\|_{*}$.

C.2 Details of implementing the max flow algorithm

The max flow problem requires several inputs, namely vertices V, arcs E, source s, sink n. In addition, each arc has flow f, capacity c, and cost y . To solve the inverse projection step in the algorithm, we need to translate those elements into our context.

Define a node as a single variable ($\boldsymbol{x}_j \in \mathbb{V}$) or a set of variables ($\mathbb{g}_k \in \mathbb{G}$). Let V be $\mathbb{V} \cup \mathbb{G}$, and $E \subseteq V \times V$. E contains three types of arcs. We show them with their flow, capacity and flow below.

Type	**From**	**To**	**Flow** f	**Capacity** c	**Cost** y
1	s	$\mathbb{g}_k \in \mathbb{G}$	$\sum_{\boldsymbol{x}_j \in \mathbb{V}} \sum_{\mathbb{g} \in \mathbb{G}} \boldsymbol{\xi}^{j}_{\mid \mathbb{g}}$	$\lambda\omega_{\mathbb{g}}$	0
2	$\mathbb{g}_k$	$\boldsymbol{x}_j \in \mathbb{g}_k$	$\boldsymbol{\xi}^{j}_{\mid \mathbb{g}}$	∞	0
3	$\boldsymbol{x}_j \in \mathbb{V}$	n	$(\sum_{\mathbb{g} \in \mathbb{G}} \boldsymbol{\xi}^{j}_{\mid \mathbb{g}})_{\boldsymbol{x}_j \in \mathbb{V}}$	∞	M^{*}

$^{*}M = \frac{1}{2t}[\beta^{k}_{j} - t\nabla f(\beta^{k}_{j}) - (\sum_{\mathbb{g} \in \mathbb{G}} \boldsymbol{\xi}^{j}_{\mid \mathbb{g}})_{\boldsymbol{x}_j \in \mathbb{V}}]^{2}$

Table C.1: Corresponding inputs of the max flow algorithm

After having the corresponding input, we can implement the max flow algorithm, and thus proceed with the inverse projection step, i.e., updating $(\sum_{\mathbb{g} \in \mathbb{G}} \boldsymbol{\xi}^{j}_{\mid \mathbb{g}})_{\boldsymbol{x}_j \in \mathbb{V}}$ by finding its maximum value with the minimum cost while satisfying relevant constraints: each flow is smaller than the capacity.

C.3 Grouping structure identification

Overlapping group Lasso can enforce a number of groups of variable coefficients to be 0 with a certain level of penalization. The remaining variables are said being selected.

Consider the selection dependency "if $\{A_1B, A_2B\}$ is selected, then $\{A_1, A_2, B\}$ must be selected". Suppose for now all candidate variables are $\mathbb{V} = \{A_1, A_2, B, A_1B, A_2B\}$, which are the variables that are involved in this rule. According to Table 3.2 in Chapter 3, the selection dictionary (all permissible subsets of covariates that respect the selection dependency) is $\mathbb{D} = \{\emptyset, \{A_1, A_2\}, \{B\}, \{A_1, A_2, B\}, \{A_1, A_2, B, A_1B, A_2B\}\}$. Based on Theorem 5 in Chapter 3, we need to create groups whose complements (and their combinations) are equal to $\mathbb{D} \setminus \mathbb{V}$. We thus postulate three groups: $\{A_1, A_2, A_1B, A_2B\}$, $\{B, A_1B, A_2B\}$ and $\{A_1B, A_2B\}$, which satisfy the requirement. Similarly, to respect the selection dependency "if $\{C_1B, C_2B\}$ is selected, then $\{C_1, C_2, B\}$ must be selected", we postulate another three groups $\{C_1, C_2$, C_1B , $C_2B\}$, $\{B, C_1B, C_2B\}$ and $\{C_1B, C_2B\}$.

However, the two dependencies share a same variable B: if either $\{A_1B, A_2B\}$ or $\{C_1B, C_2B\}$ is selected, then B must be selected. To satisfy this requirement, we need to merge the two groups $\{B, A_1B, A_2B\}$ and $\{B, C_1B, C_2B\}$ into one group $\{A_1B, A_2B, B, C_1B, C_2B\}$ to prevent the occurrence of rule-breaking combinations for example, $\{C_1B, C_2B\}$ being selected without B.

We also need to respect another selection dependency: the dummy variables for a categorical variable need to be selected collectively. The categorical interaction variables AB and BC are already being selected collectively because of the above selection dependencies. However, additional groups for $\{A_1, A_2\}$ are unnecessary as this would make it possible to select $\{A_1B, A_2B\}$ without A. In addition, with the above groups, A_1, and A_2 would never be selected individually because they are always in a same group.

Therefore, we have 5 defined groups listed below

$$\mathfrak{g}_1 = \{A_1, A_2, A_1B, A_2B\}, \mathfrak{g}_2 = \{B, A_1B, A_2B, C_1B, C_2B\},$$

$$\mathfrak{g}_3 = \{A_1B, A_2B\}, \mathfrak{g}_4 = \{C_1, C_2, C_1B, C_2B\}, \mathfrak{g}_5 = \{C_1B, C_2B\}.$$

References

Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer, 1998.

Pontus Andell, Sasha Koul, Andreas Martinsson, Johan Sundström, Tomas Jernberg, J Gustav Smith, Stefan James, Bertil Lindahl, and David Erlinge. Impact of chronic obstructive pulmonary disease on morbidity and mortality after myocardial infarction. *Open heart*, 1 (1):e000002, 2014.

Charlotte Møller Andersen and Rasmus Bro. Variable selection in regression—a tutorial. *Journal of chemometrics*, 24(11-12):728–737, 2010.

Maxim Babenko and Andrew V Goldberg. Experimental evaluation of a parametric flow algorithm. *Technical Report MSR-TR-2006-77*, 2006.

Francis Bach. Structured sparsity-inducing norms through submodular functions. *Advances in Neural Information Processing Systems*, 23, 2010.

Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Structured sparsity through convex optimization. *Statistical Science*, 27(4):450–468, 2012a.

Francis Bach, Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, et al. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1): 1–106, 2012b.

Soudabeh Khojasteh Banankhah, Erika Friedmann, and Sue Thomas. Effective treatment of depression improves post-myocardial infarction survival. *World journal of cardiology*, 7 (4):215, 2015.

Harry Bartelink, Jean-Claude Horiot, Philip Poortmans, Henk Struikmans, Walter Van den Bogaert, Isabelle Barillot, Alain Fourquet, Jacques Borger, Jos Jager, Willem Hoogenraad, et al. Recurrence rates after treatment of breast cancer with standard radiotherapy with or without additional radiation. *New England Journal of Medicine*, 345(19):1378–1387, 2001.

Thomas Lumley based on Fortran code by Alan Miller. *leaps: Regression Subset Selection*, 2020. URL https://CRAN.R-project.org/package=leaps. R package version 3.1.

Amir Beck. *First-order methods in optimization*, volume 25. SIAM, 2017.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Stephen Becker, Jérôme Bobin, and Emmanuel J Candès. Nesta: A fast and accurate first-order method for sparse recovery. *SIAM Journal on Imaging Sciences*, 4(1):1–39, 2011.

Axel Benner, Manuela Zucknick, Thomas Hielscher, Carina Ittrich, and Ulrich Mansmann. High-dimensional cox models: the choice of penalty as part of the model building process. *Biometrical Journal*, 52(1):50–69, 2010.

Dimitri Bertsekas. *Network optimization: continuous and discrete models*, volume 8. Athena Scientific, 1998.

Dimitri P Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334, 1997.

Dimitri P Bertsekas, W Hager, and O Mangasarian. Nonlinear programming. athena scientific belmont. *Massachusets, USA*, 1999.

Dimitris Bertsimas and Angela King. Logistic regression: From art to science. *Statistical Science*, pages 367–384, 2017.

Dimitris Bertsimas and Robert Weismantel. *Optimization over integers*, volume 13. Dynamic Ideas Belmont, 2005.

Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. *The annals of statistics*, 44(2):813–852, 2016.

Sahir R Bhatnagar, Amanda Lovato, Yi Yang, and Celia MT Greenwood. Sparse additive interaction learning. *bioRxiv*, 2018. 10.1101/445304. URL https://www.biorxiv.org/content/early/2018/10/16/445304.

Sahir R Bhatnagar, Tianyuan Lu, Amanda Lovato, David L Olds, Michael S Kobor, Michael J Meaney, Kieran O'Donnell, Yi Yang, and Celia MT Greenwood. A sparse additive model for high-dimensional interactions with an exposure variable. *BioRxiv*, page 445304, 2020.

Jacob Bien, Jonathan Taylor, and Robert Tibshirani. A lasso for hierarchical interactions. *Annals of statistics*, 41(3):1111, 2013.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Occam's razor. *Information processing letters*, 24(6):377–380, 1987.

Jonathan Borwein and Adrian S Lewis. *Convex analysis and nonlinear optimization: theory and examples*. Springer Science & Business Media, 2010.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Jelena Bradic, Jianqing Fan, and Jiancheng Jiang. Regularization for cox's proportional hazards model with np-dimensionality. *Annals of statistics*, 39(6):3092, 2011.

Harold J Breaux. On stepwise multiple linear regression. Technical report, Army Ballistic Research Lab Aberdeen Proving Ground MD, 1967.

Patrick Breheny and Jian Huang. Penalized methods for bi-level variable selection. *Statistics and its interface*, 2(3):369, 2009.

Patrick Breheny and Jian Huang. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The annals of applied statistics*, 5(1):232, 2011.

Patrick Breheny and Jian Huang. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and computing*, 25(2): 173–187, 2015.

Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. *Classification and regression trees*. Routledge, 2017.

Norman Breslow. Covariance analysis of censored survival data. *Biometrics*, pages 89–99, 1974.

John Burn and Munir Pirmohamed. Direct oral anticoagulants versus warfarin: is new always better than the old? *Open Heart*, 5(1), 2018.

Frederick Campbell and Genevera I Allen. Within group variable selection through the exclusive lasso. *Electronic Journal of Statistics*, 11(2):4220–4257, 2017.

Tze-Fan Chao, Gregory YH Lip, Yenn-Jiang Lin, Shih-Lin Chang, Li-Wei Lo, Yu-Feng Hu, Ta-Chuan Tuan, Jo-Nan Liao, Fa-Po Chung, Tzeng-Ji Chen, et al. Major bleeding and intracranial hemorrhage risk prediction in patients with atrial fibrillation: attention to modifiable bleeding risk factors or use of a bleeding risk stratification score? a nationwide cohort study. *International journal of cardiology*, 254:157–161, 2018.

Yuchen Chen and Yuhong Yang. The one standard error rule for model selection: Does it work? *Stats*, 4(4):868–892, 2021.

Boris V Cherkassky and Andrew V Goldberg. On implementing the push—relabel method for the maximum flow problem. *Algorithmica*, 19(4):390–410, 1997.

Hugh Chipman. Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, 24(1):17–36, 1996.

Nam Hee Choi, William Li, and Ji Zhu. Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association*, 105(489):354–364, 2010.

Mohammad Ziaul Islam Chowdhury and Tanvir C Turin. Variable selection strategies and its importance in clinical prediction modelling. *Family medicine and community health*, 8 (1), 2020.

E Christensen, P Schlichting, P Kragh Andersen, L Fauerholdt, G Schou, B Vestergaard Pedersen, E Juhl, H Poulsen, N Tygstrup, and Copenhagen Study Group for Liver Diseases. Updating prognosis and therapeutic effect evaluation in cirrhosis with cox's multiple regression model for time-dependent variables. *Scandinavian journal of gastroenterology*, 21(2):163–174, 1986.

Yeonseung Chung and David B Dunson. Nonparametric bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association*, 104(488): 1646–1660, 2009.

J'Neka S Claxton, Richard F MacLehose, Pamela L Lutsey, Faye L Norby, Lin Y Chen, Wesley T O'Neal, Alanna M Chamberlain, Lindsay GS Bengtson, and Alvaro Alonso. A new model to predict major bleeding in patients with atrial fibrillation using warfarin or direct oral anticoagulants. *PloS one*, 13(9):e0203599, 2018.

Mario Cleves, William Gould, William W Gould, Roberto Gutierrez, and Yulia Marchenko. *An introduction to survival analysis using Stata*. Stata press, 2008.

Patrick L Combettes and Jean-Christophe Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer, 2011.

SJ Connolly, J Pogue, J Eikelboom, ACTIVE W Investigators, et al. Benefit of oral anticoagulant over antiplatelet therapy in af depends on the quality of the inr control achieved as measured by time in therapeutic range. *Circulation*, 118:2029, 2008.

David R Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232, 1958.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Mattheus HJ de Bruijne, Saskia le Cessie, Hanneke C Kluin-Nelemans, and Hans C van Houwelingen. On the use of cox regression in the presence of an irregularly observed time-dependent covariate. *Statistics in medicine*, 20(24):3817–3829, 2001.

Jurgen A Doornik. Econometric model selection with more variables than observations. Technical report, Citeseer, 2009.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

Michael Alin Efroymson. Multiple regression analysis. *Mathematical methods for digital computers*, pages 191–203, 1960.

Tewodros Eguale, Nancy Winslade, James A Hanley, David L Buckeridge, and Robyn Tamblyn. Enhancing pharmacosurveillance with systematic collection of treatment indication in electronic prescribing. *Drug safety*, 33(7):559–567, 2010.

Robert C Elston, Jane M Olson, and Lyle Palmer. *Biostatistical genetics and genetic epidemiology*, volume 1. John Wiley & Sons, 2002.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

Jianqing Fan and Jinchi Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101, 2010.

Lloyd D Fisher and Danyu Y Lin. Time-dependent covariates in the cox proportional-hazards regression model. *Annual review of public health*, 20(1):145–157, 1999.

Julien MAIRAL for C++ SPAMS code, with contribution from Jean-Paul CHIEZE for interface to language R, and Ghislain DURIF for the porting to R-3.x. *spams: R interface to some functions of library spams*, 2017. R package version 2.6.

Lester Randolph Ford and Delbert R Fulkerson. Maximal flow through a network. *Canadian journal of Mathematics*, 8:399–404, 1956.

LLdiko E Frank and Jerome H Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

Leif Friberg, Mårten Rosenqvist, and Gregory YH Lip. Net clinical benefit of warfarin in patients with atrial fibrillation: a report from the swedish atrial fibrillation cohort study. *Circulation*, 125(19):2298–2307, 2012.

Jerome Friedman, Trevor Hastie, Holger Höfling, and Robert Tibshirani. Pathwise coordinate optimization. *The annals of applied statistics*, 1(2):302–332, 2007.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010a.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*, 2010b.

Giorgio Gallo, Michael D Grigoriadis, and Robert E Tarjan. A fast parametric maximum flow algorithm and applications. *SIAM Journal on Computing*, 18(1):30–55, 1989.

Svetlana V Garkina, Tatiana V Vavilova, Dmitry S Lebedev, and Evgeny N Mikhaylov. Compliance and adherence to oral anticoagulation therapy in elderly patients with atrial fibrillation in the era of direct oral anticoagulants. *Journal of geriatric cardiology: JGC*, 13(9):807, 2016.

H Gauch Jr. Parsimony and efficiency. *Scientific Method in Practice*, pages 269–326, 2002.

Robin Genuer, Jean-Michel Poggi, and Christine Tuleau-Malot. Variable selection using random forests. *Pattern recognition letters*, 31(14):2225–2236, 2010.

Edward I George. The variable selection problem. *Journal of the American Statistical Association*, 95(452):1304–1308, 2000.

Alan S Go, Daniel E Singer, Sengwee Toh, T Craig Cheetham, Marsha E Reichman, David J Graham, Mary Ross Southworth, Rongmei Zhang, Rima Izem, Margie R Goulding, et al. Outcomes of dabigatran and warfarin for atrial fibrillation in contemporary practice: a retrospective cohort study. *Annals of internal medicine*, 167(12):845–854, 2017.

Andrew V Goldberg and Robert E Tarjan. A new approach to the maximum-flow problem. *Journal of the ACM (JACM)*, 35(4):921–940, 1988.

David J Graham, Marsha E Reichman, Michael Wernecke, Rongmei Zhang, Mary Ross Southworth, Mark Levenson, Ting-Chang Sheu, Katrina Mott, Margie R Goulding, Monika Houstoun, et al. Cardiovascular, bleeding, and mortality risks in elderly medicare patients treated with dabigatran or warfarin for nonvalvular atrial fibrillation. *Circulation*, 131(2):157–164, 2015.

Sander Greenland. Modeling and variable selection in epidemiologic analysis. *American journal of public health*, 79(3):340–349, 1989.

David Gregoratti, Xavier Mestre, and Carlos Buelga. Exclusive group lasso for structured variable selection. *arXiv preprint arXiv:2108.10284*, 2021.

Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2022. URL https://www.gurobi.com.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

Michael Hamada and CF Jeff Wu. Analysis of designed experiments with complex aliasing. *Journal of quality technology*, 24(3):130–137, 1992.

Elizabeth Handorf, Yinuo Yin, Michael Slifker, and Shannon Lynch. Variable selection in social-environmental data: sparse regression and tree ensemble machine learning approaches. *BMC Medical Research Methodology*, 20(1):1–10, 2020.

Edward J Hannan and Barry G Quinn. The determination of the order of an autoregression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):190–195, 1979.

Asad Haris, Daniela Witten, and Noah Simon. Convex modeling of interactions with strong heredity. *Journal of Computational and Graphical Statistics*, 25(4):981–1004, Oct 2016a. ISSN 1537-2715. 10.1080/10618600.2015.1067217. URL http://dx.doi.org/10.1080/10618600.2015.1067217.

Asad Haris, Daniela Witten, and Noah Simon. Convex modeling of interactions with strong heredity. *Journal of Computational and Graphical Statistics*, 25(4):981–1004, 2016b.

Frank E Harrell Jr, Kerry L Lee, Robert M Califf, David B Pryor, and Robert A Rosati. Regression modelling strategies for improved prognostic prediction. *Statistics in medicine*, 3(2):143–152, 1984.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Georg Heinze, Christine Wallisch, and Daniela Dunkler. Variable selection–a review and recommendations for the practicing statistician. *Biometrical journal*, 60(3):431–449, 2018.

Dorit S Hochbaum and Sung-Pil Hong. About strongly polynomial time algorithms for quadratic optimization over submodular constraints. *Mathematical programming*, 69(1): 269–309, 1995.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Jian Huang, Shuange Ma, Huiliang Xie, and Cun-Hui Zhang. A group bridge approach for variable selection. *Biometrika*, 96(2):339–355, 2009.

Jian Huang, Joel L Horowitz, and Fengrong Wei. Variable selection in nonparametric additive models. *Annals of statistics*, 38(4):2282, 2010.

Jian Huang, Patrick Breheny, and Shuangge Ma. A selective review of group selection in high-dimensional models. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 27(4), 2012.

Junzhou Huang and Tong Zhang. The benefit of group sparsity. *The Annals of Statistics*, 38(4):1978–2004, 2010.

Junzhou Huang, Tong Zhang, and Dimitris Metaxas. Learning with structured sparsity. *Journal of Machine Learning Research*, 12(11), 2011.

Clifford M Hurvich and Chih-Ling Tsai. Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307, 1989.

Clifford M Hurvich and Chih—Ling Tsai. The impact of model selection on inference in linear regression. *The American Statistician*, 44(3):214–217, 1990.

Laurent Jacob, Guillaume Obozinski, and Jean-Philippe Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th annual international conference on machine learning*, pages 433–440, 2009.

Rodolphe Jenatton, Jean-Yves Audibert, and Francis Bach. Structured variable selection with sparsity-inducing norms. *Journal of Machine Learning Research*, 12(Oct):2777–2824, 2011a.

Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, and Francis Bach. Proximal methods for hierarchical sparse coding. *Journal of Machine Learning Research*, 12(Jul):2297–2334, 2011b.

Rodolphe Jenatton, Alexandre Gramfort, Vincent Michel, Guillaume Obozinski, Evelyn Eger, Francis Bach, and Bertrand Thirion. Multiscale mining of fmri data with hierarchical structured sparsity. *SIAM Journal on Imaging Sciences*, 5(3):835–856, 2012.

Jerald B Johnson and Kristian S Omland. Model selection in ecology and evolution. *Trends in ecology & evolution*, 19(2):101–108, 2004.

V Roshan Joseph. A bayesian approach to the design and analysis of fractionated experiments. *Technometrics*, 48(2):219–229, 2006.

John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011.

Markus Kalisch and Peter Bühlman. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(3), 2007.

Seyoung Kim and Eric P. Xing. Tree-guided group lasso for multi-task regression with structured sparsity. In Johannes Fürnkranz and Thorsten Joachims, editors, *Proceedings*

of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel, pages 543–550. Omnipress, 2010. URL https://icml.cc/Conferences/2010/papers/352.pdf.

Seyoung Kim and Eric P Xing. Tree-guided group lasso for multi-response regression with structured sparsity, with an application to eqtl mapping. *The Annals of Applied Statistics*, 6(3):1095–1117, 2012.

Deguang Kong, Ryohei Fujimaki, Ji Liu, Feiping Nie, and Chris Ding. Exclusive feature learning on arbitrary structures via $\ell_{\{}1, 2\}$-norm. *Advances in neural information processing systems*, 27, 2014.

Lynn Kuo and Bani Mallick. Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 65–81, 1998.

Simon Lacoste-Julien, Fei Sha, and Michael Jordan. Disclda: Discriminative learning for dimensionality reduction and classification. *Advances in neural information processing systems*, 21, 2008.

C Seth Landefeld and Ohio Lee Goldman. Major bleeding in outpatients treated with warfarin: incidence and prediction by factors known at the start of outpatient therapy. *The American journal of medicine*, 87(2):144–152, 1989.

Deirdre A Lane and Gregory YH Lip. Use of the cha2ds2-vasc and has-bled scores to aid decision making for thromboprophylaxis in nonvalvular atrial fibrillation. *Circulation*, 126 (7):860–865, 2012.

Kenneth Lange, David R Hunter, and Ilsoon Yang. Optimization transfer using surrogate objective functions. *Journal of computational and graphical statistics*, 9(1):1–20, 2000.

Julie C Lauffenburger, Joel F Farley, Anil K Gehi, Denise H Rhoney, M Alan Brookhart, and Gang Fang. Effectiveness and safety of dabigatran and warfarin in real-world us patients

with non-valvular atrial fibrillation: a retrospective cohort study. *Journal of the American Heart Association*, 4(4):e001798, 2015.

Rafael Lazimy. Mixed-integer quadratic programming. *Mathematical Programming*, 22(1): 332–349, 1982.

Michael Lim and Trevor Hastie. Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics*, 24(3):627–654, 2015.

Meixia Lin, Defeng Sun, Kim-Chuan Toh, and Yancheng Yuan. A dual newton based preconditioned proximal point algorithm for exclusive lasso models. *arXiv preprint arXiv:1902.00151*, 2019.

Heinz Linhart and Walter Zucchini. *Model selection.* John Wiley & Sons, 1986.

Gregory YH Lip and Hung-Fat Tse. Management of atrial fibrillation. *The Lancet*, 370 (9587):604–618, 2007.

Shuangge Ma, Xiao Song, and Jian Huang. Supervised group lasso with applications to microarray data analysis. *BMC bioinformatics*, 8(1):1–17, 2007.

Julien Mairal, Rodolphe Jenatton, Guillaume Obozinski, and Francis Bach. Network flow algorithms for structured sparsity. *arXiv preprint arXiv:1008.5209*, 2010.

Julien Mairal, Francis Bach, and Jean Ponce. Sparse modeling for image and vision processing. *arXiv preprint arXiv:1411.3230*, 2014.

Colin L Mallows. Some comments on cp. *Technometrics*, 42(1):87–94, 2000.

Géric Maura, Pierre-Olivier Blotière, Kim Bouillon, Cécile Billionnet, Philippe Ricordeau, François Alla, and Mahmoud Zureik. Comparison of the short-term risk of bleeding and arterial thromboembolic events in nonvalvular atrial fibrillation patients newly treated with dabigatran or rivaroxaban versus vitamin k antagonists: a french nationwide propensity-matched cohort study. *Circulation*, 132(13):1252–1260, 2015.

Rahul Mazumder, Jerome H Friedman, and Trevor Hastie. Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 106(495):1125–1138, 2011.

Tahir Mehmood, Kristian Hovde Liland, Lars Snipen, and Solve Sæbø. A review of variable selection methods in partial least squares regression. *Chemometrics and intelligent laboratory systems*, 118:62–69, 2012.

Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

Nicolai Meinshausen and Peter Bühlmann. Variable selection and high-dimensional graphs with the lasso. *Ann Stat*, 34:1436–1462, 2006.

Jean Jacques Moreau. Fonctions convexes duales et points proximaux dans un espace hilbertien. *Comptes rendus hebdomadaires des séances de l'Académie des sciences*, 255:2897–2899, 1962.

Yu. Nesterov. Gradient methods for minimizing composite objective function. CORE Discussion Papers 2007076, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2007. URL https://EconPapers.repec.org/RePEc:cor:louvco:2007076.

Yu Nesterov. Gradient methods for minimizing composite functions. *Mathematical programming*, 140(1):125–161, 2013a.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013b.

Kehinde O Obamiro, Leanne Chalmers, and Luke RE Bereznicki. A summary of the literature evaluating adherence and persistence with oral anticoagulants in atrial fibrillation. *American Journal of Cardiovascular Drugs*, 16(5):349–363, 2016.

Guillaume Obozinski, Ben Taskar, and Michael I Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, 20(2): 231–252, 2010.

Guillaume Obozinski, Laurent Jacob, and Jean-Philippe Vert. Group lasso with overlaps: the latent group lasso approach. *arXiv preprint arXiv:1110.0413*, 2011a.

Guillaume Obozinski, Laurent Jacob, and Jean-Philippe Vert. Group lasso with overlaps: the latent group lasso approach. *arXiv preprint arXiv:1110.0413*, 2011b.

Fernanda Araujo Pimentel Peres and Flavio Sanson Fogliatto. Variable selection methods in multivariate statistical process control: A systematic literature review. *Computers & Industrial Engineering*, 115:603–619, 2018.

Sylvie Perreault, Robert Côté, Brian White-Guay, Marc Dorais, Essaïd Oussaïd, and Mireille E Schnitzer. Anticoagulants in older patients with nonvalvular atrial fibrillation after intracranial hemorrhage. *Journal of stroke*, 21(2):195, 2019.

Sylvie Perreault, Simon de Denus, Brian White-Guay, Robert Côté, Mireille E Schnitzer, Marie-Pierre Dubé, Marc Dorais, and Jean-Claude Tardif. Oral anticoagulant prescription trends, profile use, and determinants of adherence in patients with atrial fibrillation. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, 40(1):40–54, 2020.

Ron Pisters, Deirdre A Lane, Robby Nieuwlaat, Cees B De Vos, Harry JGM Crijns, and Gregory YH Lip. A novel user-friendly score (has-bled) to assess 1-year risk of major bleeding in patients with atrial fibrillation: the euro heart survey. *Chest*, 138(5):1093–1100, 2010.

Yves Pochet and Laurence A Wolsey. *Production planning by mixed integer programming*. Springer Science & Business Media, 2006.

Benjamin Poignard. Asymptotic theory of the sparse group lasso. *arXiv preprint arXiv:1611.06034*, 2016.

John N Primrose, Rafael Perera, Alastair Gray, Peter Rose, Alice Fuller, Andrea Corkhill, Steve George, David Mant, FACS Trial Investigators, et al. Effect of 3 to 5 years of scheduled cea and ct follow-up to detect recurrence of colorectal cancer: the facs randomized clinical trial. *Jama*, 311(3):263–270, 2014.

Jakub Z Qazi, Mireille E Schnitzer, Robert Côté, Marie-Josée Martel, Marc Dorais, and Sylvie Perreault. Predicting major bleeding among hospitalized patients using oral anticoagulants for atrial fibrillation after discharge. *PloS one*, 16(3):e0246691, 2021.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL https://www.R-project.org/.

Bruria Hirsh Raccah, Amihai Rottenstreich, Netanel Zacks, Mordechai Muszkat, Ilan Matok, Amichai Perlman, and Yosef Kalish. Drug interaction as a predictor of direct oral anticoagulant drug levels in atrial fibrillation patients. *Journal of thrombosis and thrombolysis*, 46(4):521–527, 2018.

Peter Radchenko and Gareth M James. Variable selection using adaptive nonlinear interaction structures in high dimensions. *Journal of the American Statistical Association*, 105 (492):1541–1553, 2010.

Juha Reunanen. Overfitting in making comparisons between variable selection methods. *Journal of Machine Learning Research*, 3(Mar):1371–1382, 2003.

Vanessa Roldán, Francisco Marín, Sergio Manzano-Fernández, Pilar Gallego, Juan Antonio Vílchez, Mariano Valdés, Vicente Vicente, and Gregory YH Lip. The has-bled score

has better prediction accuracy for major bleeding than chads2 or cha2ds2-vasc scores in anticoagulated patients with atrial fibrillation. *Journal of the American College of Cardiology*, 62(23):2199–2204, 2013.

Volker Roth and Bernd Fischer. The group-lasso for generalized linear models: uniqueness of solutions and efficient algorithms. In *Proceedings of the 25th international conference on Machine learning*, pages 848–855, 2008.

Eduardo Sabaté and Eduardo Sabaté. *Adherence to long-term therapies: evidence for action*. World Health Organization, 2003.

Mireille E Schnitzer, Robert W Platt, and Madeleine Durand. A tutorial on dealing with time-varying eligibility for treatment: Comparing the risk of major bleeding with direct-acting oral anticoagulant s vs warfarin. *Statistics in medicine*, 39(29):4538–4550, 2020.

C Mary Schooling and Heidi E Jones. Clarifying questions about "risk factors": predictors versus explanation. *Emerging themes in epidemiology*, 15(1):1–6, 2018.

Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

Mary Beth Seasholtz and Bruce Kowalski. The parsimony principle applied to multivariate calibration. *Analytica Chimica Acta*, 277(2):165–177, 1993.

Yiyuan She, Zhifeng Wang, and He Jiang. Group regularized estimation under structural hierarchy. *Journal of the American Statistical Association*, 113(521):445–454, 2018.

Galit Shmueli. To explain or to predict? *Statistical science*, 25(3):289–310, 2010.

Susan M Shortreed and Ashkan Ertefaie. Outcome-adaptive lasso: variable selection for causal inference. *Biometrics*, 73(4):1111–1122, 2017.

Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of statistical software*, 39(5):1, 2011.

Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of computational and graphical statistics*, 22(2):231–245, 2013.

Michael Smith and Robert Kohn. Nonparametric regression using bayesian variable selection. *Journal of Econometrics*, 75(2):317–343, 1996.

Ewout W Steyerberg, Marinus JC Eijkemans, and J Dik F Habbema. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. *Journal of clinical epidemiology*, 52(10):935–942, 1999.

Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society: Series B (Methodological)*, 36(2):111–133, 1974.

Hokeun Sun, Wei Lin, Rui Feng, and Hongzhe Li. Network-regularized high-dimensional cox regression for analysis of genomic data. *Statistica Sinica*, 24(3):1433, 2014.

Yuxin Sun, Benny Chain, Samuel Kaski, and John Shawe-Taylor. Correlated feature selection with extended exclusive group lasso. *arXiv preprint arXiv:2002.12460*, 2020.

Marie-Pierre Sylvestre and Michal Abrahamowicz. Comparison of algorithms to generate event times conditional on time-dependent covariates. *Statistics in medicine*, 27(14):2618–2634, 2008.

Marie-Pierre Sylvestre, Thad Evans, Todd MacKenzie, and Michal Abrahamowicz. *PermAlgo: Permutational algorith to generate event times conditional on a covariate matrix including time-dependent covariates*, 2010. R package version 1.1.

Robyn Tamblyn, Gilles Lavoie, Lina Petrella, and Johanne Monette. The use of prescription claims databases in pharmacoepidemiological research: the accuracy and comprehensive-

ness of the prescription claims database in quebec. *Journal of clinical epidemiology*, 48 (8):999–1009, 1995.

Zaixiang Tang, Shufeng Lei, Xinyan Zhang, Zixuan Yi, Boyi Guo, Jake Y Chen, Yueping Shen, and Nengjun Yi. Gsslasso cox: a bayesian hierarchical model for predicting survival and detecting associated genes by incorporating pathway information. *BMC bioinformatics*, 20(1):94, 2019.

Terry M Therneau. *A Package for Survival Analysis in R*, 2021. URL https://CRAN.R-project.org/package=survival. R package version 3.2-10.

Warren R Thompson. *Variable selection of correlated predictors in logistic regression: investigating the diet-heart hypothesis*. The Florida State University, 2009.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Robert Tibshirani. The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395, 1997.

Armando Tripodi, Simon Braham, Barbara Scimeca, Marco Moia, and Flora Peyvandi. How and when to measure anticoagulant effects of direct oral anticoagulants? practical issues. *Polish archives of internal medicine*, 128(6):379–385, 2018.

Gaël Varoquaux, Alexandre Gramfort, Fabian Pedregosa, Vincent Michel, and Bertrand Thirion. Multi-subject dictionary learning to segment an atlas of brain spontaneous activity. In *Biennial International Conference on information processing in medical imaging*, pages 562–573. Springer, 2011.

Todd C Villines, Janet Schnee, Kathy Fraeman, Kimberly Siu, Matthew W Reynolds, Jenna Collins, and Eric Schwartzman. A comparison of the safety and effectiveness of dabiga-

tran and warfarin in non-valvular atrial fibrillation patients in a large healthcare system. *Thrombosis and haemostasis*, 114(12):1290–1298, 2015.

ML Wallace. Time-dependent tree-structured survival analysis with unbiased variable selection through permutation tests. *Statistics in medicine*, 33(27):4790–4804, 2014.

Guanbo Wang, Mireille E Schnitzer, Tom Chen, Rui Wang, and Robert W Platt. A general framework for identification of permissible variable subsets in structured model selection. *arXiv preprint arXiv:2110.01031*, 2021.

Hansheng Wang and Chenlei Leng. A note on adaptive group lasso. *Computational statistics & data analysis*, 52(12):5277–5286, 2008.

Lifeng Wang, Guang Chen, and Hongzhe Li. Group scad regression analysis for microarray time course gene expression data. *Bioinformatics*, 23(12):1486–1494, 2007.

Lifeng Wang, Hongzhe Li, and Jianhua Z Huang. Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association*, 103(484):1556–1569, 2008.

Qin Wang and Xiangrong Yin. A nonlinear multi-dimensional variable selection method for high dimensional data: Sparse mave. *Computational Statistics & Data Analysis*, 52(9): 4512–4520, 2008.

Sijian Wang, B Nan, N Zhu, and J Zhu. Hierarchically penalized cox regression with grouped variables. *Biometrika*, 96(2):307–322, 2009.

Larry Wasserman and Kathryn Roeder. High dimensional variable selection. *Annals of statistics*, 37(5A):2178, 2009.

Mark J Whittingham, Philip A Stephens, Richard B Bradbury, and Robert P Freckleton. Why do we still use stepwise modelling in ecology and behaviour? *Journal of animal ecology*, 75(5):1182–1189, 2006.

Machelle Wilchesky, Robyn M Tamblyn, and Allen Huang. Validation of diagnostic codes within medical services claims. *Journal of clinical epidemiology*, 57(2):131–141, 2004.

Stephen J Wright, Robert D Nowak, and Mário AT Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7):2479–2493, 2009.

Chao Wu, Guolong Wang, Simon Hu, Yue Liu, Hong Mi, Ye Zhou, Yi-ke Guo, and Tongtong Song. A data driven methodology for social science research with left-behind children as a case study. *Plos one*, 15(11):e0242483, 2020.

Tong Tong Wu and Kenneth Lange. Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, 2(1):224–244, 2008.

Makoto Yamada, Takeuchi Koh, Tomoharu Iwata, John Shawe-Taylor, and Samuel Kaski. Localized lasso for high-dimensional regression. In *Artificial Intelligence and Statistics*, pages 325–333. PMLR, 2017.

Kazuo Yamashiro, Naohide Kurita, Ryota Tanaka, Yuji Ueno, Nobukazu Miyamoto, Kenichiro Hira, Sho Nakajima, Takao Urabe, and Nobutaka Hattori. Adequate adherence to direct oral anticoagulant is associated with reduced ischemic stroke severity in patients with atrial fibrillation. *Journal of Stroke and Cerebrovascular Diseases*, 28(6): 1773–1780, 2019.

Xiaohan Yan, Jacob Bien, et al. Hierarchical sparse modeling: A choice of two group lasso formulations. *Statistical Science*, 32(4):531–560, 2017.

Yi Yang and Hui Zou. A fast unified algorithm for solving group-lasso penalize learning problems. *Statistics and Computing*, 25(6):1129–1141, 2015.

Xiaoxi Yao, Neena S Abraham, Lindsey R Sangaralingham, M Fernanda Bellolio, Robert D McBane, Nilay D Shah, and Peter A Noseworthy. Effectiveness and safety of dabigatran,

rivaroxaban, and apixaban versus warfarin in nonvalvular atrial fibrillation. *Journal of the American Heart Association*, 5(6):e003725, 2016.

Tal Yarkoni and Jacob Westfall. Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6):1100–1122, 2017.

Frank Yates. *The design and analysis of factorial experiments*. Imperial Bureau of Soil Science, 1937.

Lei Yuan, Jun Liu, and Jieping Ye. Efficient methods for overlapping group lasso. *Advances in neural information processing systems*, 24:352–360, 2011.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Ming Yuan and Yi Lin. On the non-negative garrotte estimator. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):143–161, 2007.

Ming Yuan, V Roshan Joseph, and Hui Zou. Structured variable selection and estimation. *The Annals of Applied Statistics*, pages 1738–1757, 2009.

Yaohui Zeng and Patrick Breheny. Overlapping group logistic regression with applications to genetic pathway selection. *Cancer Informatics*, 15(1):179–187, 2016.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010.

Hao Helen Zhang and Wenbin Lu. Adaptive lasso for cox's proportional hazards model. *Biometrika*, 94(3):691–703, 2007.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

Peng Zhao, Guilherme Rocha, Bin Yu, et al. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37(6A):3468–3497, 2009.

Xiaowei Zhou, Can Yang, and Weichuan Yu. Moving object detection by detecting contiguous outliers in the low-rank representation. *IEEE transactions on pattern analysis and machine intelligence*, 35(3):597–610, 2012.

Yang Zhou, Rong Jin, and Steven Chu-Hong Hoi. Exclusive lasso for multi-task feature selection. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 988–995. JMLR Workshop and Conference Proceedings, 2010.

Jun Zhu, Amr Ahmed, and Eric P Xing. Medlda: maximum margin supervised topic models. *the Journal of machine Learning research*, 13(1):2237–2278, 2012.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

Hui Zou and Runze Li. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of statistics*, 36(4):1509, 2008.

Walter Zucchini. An introduction to model selection. *Journal of mathematical psychology*, 44(1):41–61, 2000.